In-Context Learning by Linear Attention: Exact Asymptotics and Experiments

Yue M. Lu[a], **Mary Letey**[a], **Jacob Zavatone-Veth**[b,c,*], **Anindita Maiti**[d,*], **Cengiz Pehlevan**[a,b,e]

[a]The John A. Paulson School of Engineering and Applied Sciences, Harvard University
[b]Center for Brain Science, Harvard University [c]Society of Fellows, Harvard University
[d]Perimeter Institute for Theoretical Physics
[e]The Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University
yuelu@seas.harvard.edu, maryletey@fas.harvard.edu, jzavatoneveth@fas.harvard.edu,
amaiti@perimeterinstitute.ca, cpehlevan@seas.harvard.edu

Abstract

Transformers have a remarkable ability to learn and execute tasks based on examples provided within the input itself, without explicit prior training. It has been argued that this capability, known as in-context learning (ICL), is a cornerstone of Transformers' success, yet questions about the necessary sample complexity and pretraining task diversity for successful ICL remain unresolved. In this work, we provide precise answers to these questions using a solvable model of ICL for a linear regression task with linear attention. We derive asymptotics for the learning curve in a regime where token dimension, context length, and pretraining diversity scale proportionally, and pretraining examples scale quadratically. Our analysis reveals a double-descent learning curve and a learning transition between low and high task diversity, which is empirically validated with experiments on realistic Transformer architectures.

1 Introduction

Since their introduction by Vaswani et al. in 2017 [1], Transformers have become a cornerstone of modern artificial intelligence (AI). Transformers achieve state-of-the art performance across many domains, even those that are not inherently sequential [2] as originally intended. Strikingly, they underpin breakthroughs achieved by large language models (LLMs) such as BERT [3], LLaMA [4], and the GPT series [5–8]. The advancements enabled by Transformers have inspired much research aimed at understanding their working principles. One key observation is that LLMs gain new behaviors and skills as their number of parameters and the size of their training datasets grow [7, 9–11]. A particularly important emergent skill is *in-context learning* (ICL), which describes the model's ability to learn and execute tasks based on the context provided within the input itself, without the need for explicit prior training on those specific tasks. ICL enables language models to perform new, specialized tasks without retraining, which is arguably a key reason for their general-purpose abilities.

Despite many recent studies on understanding ICL, important questions about how and when ICL emerges in LLMs are still mostly open. LLMs are trained (or pretrained) with a next token prediction objective. How do the different algorithmic and hyperparameter choices that go into the pretraining procedure affect ICL performance? What algorithms do Transformers implement for ICL? How many pretraining examples are required for ICL to emerge? How many examples should be provided within the input for the model to be able to solve an in-context task? How diverse should the tasks

[*]J.Z-V. and A.M. contributed equally to this work.

Mathematics of Modern Machine Learning Workshop at NeurIPS 2024.

in the training dataset be for in-context learning of truly new tasks not encountered in the training dataset? We address these questions by investigating a simplified model of a Transformer that captures its key architectural motif: the linear self-attention module [12–17]. Linear attention includes the quadratic pairwise interactions between inputs that lie at the heart of softmax attention, but it omits the normalization steps and fully connected layers. This simplification makes the model more amenable to theoretical analysis. Our main result is a sharp asymptotic analysis of ICL for linear regression using linear attention, leading to a more precisely predictive theory than previous population risk analyses or finite-sample bounds [13, 16]. The main contributions of our paper are structured as follows:

We begin in §2 by developing a simplified parameterization of linear self-attention that allows pretraining on the ICL linear regression task to be performed using ridge regression. Within this simplified model, we identify a phenomenologically rich scaling limit in which the ICL performance can be analyzed (§3). In this joint limit, we compute sharp asymptotics for ICL performance using random matrix theory. Our theoretical results reveal several interesting phenomena (§4). First, we observe double-descent in the model's ICL generalization performance as a function of pretraining dataset size, reflecting our assumption that it is pretrained to interpolation. Second, we uncover a transition to in-context learning as the pretraining task diversity increases. This transition recapitulates and builds on the empirical findings of [18] in full Transformer models. We further show through numerical experiments that these insights from our theory transfer to full Transformer models with softmax self-attention.

Understanding the mechanistic underpinnings of ICL of well-controlled synthetic tasks in solvable models is an important prerequisite to understanding how it emerges from pretraining on natural data [19].

2 Problem formulation

ICL of linear regression. In an ICL task, the model takes as input a sequence of tokens $\{x_1, y_1, x_2, y_2, \ldots, x_\ell, y_\ell, x_{\ell+1}\}$, and outputs a prediction of $y_{\ell+1}$. We will often refer to an input sequence as a *context*. We will refer to ℓ as the *context length*. We focus on an approximately linear mapping between $x_i \in \mathbb{R}^d$ and $y_i = \langle x_i, w\rangle + \epsilon_i \in \mathbb{R}$ where ϵ_i is a Gaussian noise with mean zero and variance ρ, and $w \in \mathbb{R}^d$ is referred to as a *task vector*. We note that the task vector w is fixed within a context, but can change between different contexts. The model has to learn w from the ℓ pairs presented within the context, and use it to predict $y_{\ell+1}$ from $x_{\ell+1}$.

Linear self-attention. The model that we will analytically study is the linear self-attention block [20]. Linear self-attention takes as input an embedding matrix Z, whose columns hold the sequence tokens. The choice of embedding matrix for a sequence is not unique; here, following the convention in [15, 16, 20], we will embed the input sequence $\{x_1, y_1, x_2, y_2, \ldots, x_\ell, y_\ell, x_{\ell+1}\}$ as:

$$Z = \begin{bmatrix} x_1 & x_2 & \ldots & x_\ell & x_{\ell+1} \\ y_1 & y_2 & \ldots & y_\ell & 0 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(\ell+1)}, \tag{1}$$

where 0 in the lower-right corner is a token that prompts the missing value $y_{\ell+1}$ to be predicted. For appropriately-sized key, query, and value matrices K, Q, V, the output of a linear-attention block [20–22] is given by

$$A \coloneqq Z + \frac{1}{\ell} V Z (KZ)^\top (QZ). \tag{2}$$

The output A is a matrix while our goal is to predict a scalar, $y_{\ell+1}$. Following the choice of positional encoding in (1), we will take $A_{d+1,\ell+1}$, the element of A corresponding to the 0 prompt, as the prediction for $y_{\ell+1}$, namely $\hat{y} \coloneqq A_{d+1,\ell+1}$.

Pretraining data. The model is pretrained on n sample sequences, where the μth sample is a collection of $\ell+1$ vector-scalar pairs $\{x_i^\mu \in \mathbb{R}^d, y_i^\mu \in \mathbb{R}\}_{i=1}^{\ell+1}$ related by the approximate linear mapping $y_i^\mu = \langle x_i^\mu, w^\mu\rangle + \epsilon_i^\mu$. Here, w^μ denotes the task vector associated with the μth sample. We make the following assumptions; we denote a sample from this distribution by $(Z, y_{\ell+1}) \sim \mathcal{P}_{\text{train}}$.

- x_i^μ are d-dimensional random vectors, sampled i.i.d. over both i and μ from $\mathcal{N}(0, I_d/d)$.

- At the start of training, construct a finite set of k elements, written $\Omega_k = \{w_1, w_2, \ldots, w_k\}$. The elements of this set are independently drawn once from $w_i \sim_{\text{i.i.d.}} \mathcal{N}(0, I_d)$. For $1 \le \mu \le n$, the task vector w^μ associated with the μth sample context is uniformly sampled from Ω_k. Note that the variable k controls the task diversity in the pretraining dataset. Importantly, k can be less than n, in which case the same task vector from Ω_k may be repeated multiple times.
- The noise terms ϵ_i^μ are i.i.d. over both i and μ, and drawn from $\mathcal{N}(0, \rho)$.

Parameter reduction. Before specifying a training procedure, we examine the prediction mechanism of the linear attention module for the ICL task. We start by rewriting the output of the linear attention module $\hat{y} = A_{d+1,\ell+1}$ in an alternative form. Following [16], we define

$$V = \begin{bmatrix} V_{11} & v_{12} \\ v_{21}^\top & v_{22} \end{bmatrix}, \quad M = \begin{bmatrix} M_{11} & m_{12} \\ m_{21}^\top & m_{22} \end{bmatrix} := K^\top Q, \tag{3}$$

where $V_{11} \in \mathbb{R}^{d\times d}$, $v_{12}, v_{21} \in \mathbb{R}^d$, $v_{22} \in \mathbb{R}$, $M_{11} \in \mathbb{R}^{d\times d}$, $m_{12}, m_{21} \in \mathbb{R}^d$, and $m_{22} \in \mathbb{R}$. Expanding (2), one can check that, where $\langle \cdot, \cdot \rangle$ stands for the standard inner product, we have

$$\hat{y} = \frac{1}{\ell} \left\langle x_{\ell+1}, v_{22} M_{11}^\top \sum_{i=1}^{\ell} y_i x_i + v_{22} m_{21} \sum_{i=1}^{\ell} y_i^2 + M_{11}^\top \sum_{i=1}^{\ell+1} x_i x_i^\top v_{21} + m_{21} \sum_{i=1}^{\ell} y_i x_i^\top v_{21} \right\rangle, \tag{4}$$

This expression reveals an interesting point. The first term $\frac{1}{\ell} v_{22} M_{11}^\top \sum_{i=1}^{\ell} y_i x_i$ offers a hint about how the linear attention module might be solving the task. The sum $\frac{1}{\ell} \sum_{i\le\ell} y_i x_i$ is a noisy estimate of $\mathbb{E}[xx^\top]w$ for that context. Hence, if the parameters of the model are such that $v_{22} M_{11}^\top$ is approximately $\mathbb{E}[xx^\top]^{-1}$, this term alone makes a good prediction for the output. Motivated by this observation, and a more detailed argument presented in Section SI-6 of the Supplementary Information, we study the linear attention module with the constraint $v_{21} = 0$. In this case, we have the model

$$\hat{y} = \langle \Gamma, H_Z \rangle. \tag{5}$$

for parameter matrix $\Gamma \in \mathbb{R}^{d\times(d+1)}$ and input features $H_Z \in \mathbb{R}^{d\times(d+1)}$ given by

$$\Gamma := v_{22} \begin{bmatrix} M_{11}^\top / d & m_{21} \end{bmatrix}, \qquad H_Z := x_{\ell+1} \begin{bmatrix} \frac{d}{\ell} \sum_{i\le\ell} y_i x_i^\top & \frac{1}{\ell} \sum_{i\le\ell} y_i^2 \end{bmatrix}. \tag{6}$$

Model pretraining. The parameters of the linear attention module are learned from n samples of input sequences $\{x_1^\mu, y_1^\mu, \ldots, x_{\ell+1}^\mu, y_{\ell+1}^\mu\}$ for $\mu = 1, \ldots, n$. We estimate model parameters using ridge regression, giving

$$\Gamma^* = \arg\min_{\Gamma} \sum_{\mu=1}^{n} \left(y_{\ell+1}^\mu - \langle \Gamma, H_{Z^\mu} \rangle \right)^2 + \frac{n}{d} \lambda \|\Gamma\|_{\text{F}}^2, \tag{7}$$

$$\text{vec}(\Gamma^*) = \left(\frac{n}{d} \lambda I + \sum_{\mu=1}^{n} \text{vec}(H_{Z^\mu}) \text{vec}(H_{Z^\mu})^\top \right)^{-1} \sum_{\mu=1}^{n} y_{\ell+1}^\mu \text{vec}(H_{Z^\mu}), \tag{8}$$

where $\lambda > 0$ is a regularization parameter, H_{Z^μ} refers to the input matrix (6) populated with the μth sample sequence, and $\text{vec}(\cdot)$ denotes the row-major vectorization operation.

Evaluation. For a given set of parameters Γ, the model's generalization error is defined as

$$e(\Gamma) := \mathbb{E}_{\mathcal{P}_{\text{test}}} \left[\left(y_{\ell+1} - \langle \Gamma, H_Z \rangle \right)^2 \right], \tag{9}$$

where $(Z, y_{\ell+1}) \sim \mathcal{P}_{\text{test}}$ is a new sample drawn from the probability distribution of the test dataset. We consider two different test data distributions $\mathcal{P}_{\text{test}}$:

1. *ICL task:* x_i and ϵ_i are i.i.d. Gaussians as above. However, each task vector w^{test} is drawn independently from $\mathcal{N}(0, I_d)$. We will denote the test error under this setting by $e^{\text{ICL}}(\Gamma)$.
2. *In-distribution generalization (IDG) task:* Here take $\mathcal{P}_{\text{test}} = \mathcal{P}_{\text{train}}$. In particular, the set of unique task vectors $\{w_1, \ldots, w_k\}$ is identical to that used in the pretraining data. We will denote the test error under this setting by $e^{\text{IDG}}(\Gamma)$. This task can also be referred to as *in-weight learning*.

High performance on the IDG task but low performance on the ICL task indicates that the model memorizes the training task vectors. Conversely, high performance on the ICL task suggests that the model can learn genuinely new task vectors from the provided context. To understand the performance of our model on both ICL and IDG tasks, we will need to evaluate these expressions for the pretrained attention matrix Γ^* given in (8). An asymptotically precise prediction of $e^{\mathrm{ICL}}(\Gamma^*)$ and $e^{\mathrm{IDG}}(\Gamma^*)$ will be a main result of this work.

3 Theoretical results

Joint asymptotic limit. We have now defined both the structure of the training data as well as the parameters to be optimized. For our theoretical analysis, we consider a joint asymptotic limit in which the input dimension d, the pretraining dataset size n, the context length ℓ, and the number of task vectors in the training set k, go to infinity together such that

$$\ell/d \coloneqq \alpha = \Theta(1), \quad k/d \coloneqq \kappa = \Theta(1), \quad n/d^2 \coloneqq \tau = \Theta(1). \tag{10}$$

Identification of these scalings constitutes one of the main results of our paper. As we will see, the linear attention module exhibits rich learning phenomena in this limit.

ICL and IDG learning curves. Our theoretical analysis, explained in detail in the Supplementary Information, leads to an asymptotically precise expression for the generalization error under the ICL and IDG test distributions being studied. The exact expressions of can be found in Section SI-13.2 and Section SI-13.3 of the SI. For simplicity, we only present in what follows the ridgeless limit (*i.e.*, $\lambda \to 0^+$) of the asymptotic generalization errors.

Result 1 (ICL generalization error in the ridgeless limit)**.** *Let*

$$q^* \coloneqq (1+\rho)/\alpha\,, \quad m^* \coloneqq \mathcal{M}_\kappa(q^*)\,, \quad \mu^* \coloneqq q^* \mathcal{M}_{\kappa/\tau}(q^*), \tag{11}$$

where $\mathcal{M}_\kappa(\cdot)$ is defined in (177) and $\mathcal{M}'_\kappa(\cdot)$ is the derivative of $\mathcal{M}_\kappa(q)$ with respect to q. Then

$$\lim_{\lambda\to 0^+} e^{\mathrm{ICL}}(\tau,\alpha,\kappa,\rho,\lambda) \tag{12}$$
$$= \begin{cases} \frac{\tau(1+q^*)}{1-\tau}\left[1-\tau(1-\mu^*)^2+\mu^*(\rho/q^*-1)\right] - 2\tau(1-\mu^*) + (1+\rho) & \tau < 1 \\ (q^*+1)\left(1-2q^*m^*-(q^*)^2\mathcal{M}'_\kappa(q^*)+\frac{(\rho+q^*-(q^*)^2m^*)m^*}{\tau-1}\right) - 2(1-q^*m^*) + (1+\rho) & \tau > 1 \end{cases}$$

Result 2 (IDG generalization error in the ridgeless limit)**.** *Let q^*, m^*, and μ^* be the scalars defined in (11). We have*

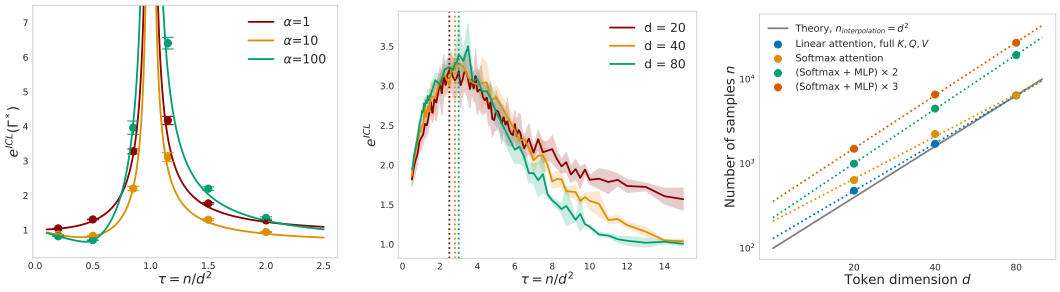
$$\lim_{\lambda\to 0^+} e^{\mathrm{IDG}}(\tau,\alpha,\kappa,\rho,\lambda) = \begin{cases} \frac{\tau}{1-\tau}\left(\frac{\rho+q^*-2q^*(1-\tau)(q^*/\xi^*+1)}{1-p^*(1-\tau)} + \frac{\tau\mu^*(q^*+\xi^*)^2}{q^*}\right) & \tau < 1 \\ \frac{\tau}{\tau-1}[\rho+q^*(1-q^*m^*)] & \tau > 1 \end{cases}, \tag{13}$$

where $\xi^* = \frac{(1-\tau)q^*}{\tau\mu^*}$ and $p^* = \left(1-\kappa\left(\frac{\kappa\xi^*}{1-\tau}+1\right)^{-2}\right)^{-1}$.

We derive this result using techniques from random matrix theory. The full setup and technical details are presented in the Supplementary Information in Section SI-9 through Section SI-13. The computations involve analysis of the properties of the finite-sample optimal parameter matrix Γ^*. We will now discuss various implications of these equations in the following sections.

4 Observed Phenomena

This section discusses two key results that are mathematically evident from our theoretical characterisation of ICL and IDG error, namely a double descent in τ and a learning transition in κ. We show how these phenomena follow directly from the theory, and further, remain present in realistic (nonlinear) transformer architectures. A detailed exposition of nonlinear architecture setup and training procedures is given in Section SI-7 in the Supplementary Info. Specific parameter configurations and more detailed descriptions of the figures are available in Section SI-8 in the Supplementary Info.

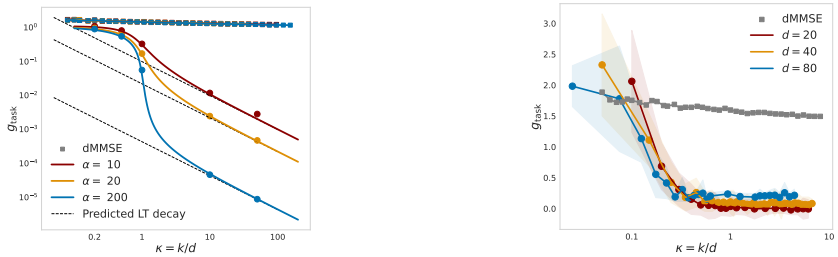


(a) Analytical error curve in τ for linear transformer

(b) Error curve in τ for nonlinear transformer

(c) Interpolation threshold follows predicted quadratic $n \propto d^2$ scaling.

Figure 1: Verification of theory in linear transformer (1a) and qualitative predictions for nonlinear models (1b,1c). Figure 1a shows theory (solid lines) vs simulations (dots). 1b shows error curves against τ for various architectures, consistent across token dimension $d = 20, 40, 80$. Double-descent phenomena is confirmed: increasing n will increase error until an interpolation threshold is reached. Coloured dashed lines indicate experimental interpolation threshold for that architecture and d configuration. Figure 1c shows that the location of the interpolation threshold occurs for n proportional to d^2 for a range of architectures, as predicted by the linear theory. Dots are experimental interpolation thresholds for various architectures, and dashed lines are best fit curves correspond to fitting $\log(n) = a\log(d) + b$, each with $a \approx 2$.



(a) g_{task} for LT and dMMSE estimator vs κ

(b) g_{task} for nonlinear transformer and dMMSE vs κ

Figure 2: Plot of $g_{\text{task}} = e^{\text{ICL}} - e^{\text{IDG}}$ against κ, illustrating sharp transition in performance as pretraining diversity increases, compared with the dMMSE estimator. Figure 2a (loglog scale) shows LT theory (solid lines) vs simulations (dots). Figure 2b shows g_{task} against κ (loglinear scale) for the nonlinear architecture given in Figure 1b, demonstrating both consistency of κ scaling across increasing dimension choices $d = 20, 40, 80$, as well as a similar sharp transition in task generalisation familiar from the linear theory.

Double-descent in pretraining samples. How large should n, the pretraining dataset size, be for the linear attention to succesfully learn the task in-context? In Figure 1a, we plot our theoretical predictions for ICL error as a function of $\tau = n/d^2$ and verify them with numerical simulations. Our results demonstrate that the quadratic scaling of sample size with input dimensions is indeed an appropriate regime where nontrivial learning phenomena can be observed.

As apparent in Figure 1a, we find that the generalization error for the ICL task is not monotonic in the number of samples. In the ridgeless limit, ICL error diverges at $\tau = 1$, with the leading order behavior proportional to $(\tau - 1)^{-1}$. This leads to a "double-descent" behavior [23, 24] in the number of samples. As in other models exhibiting double-descent [23–25], the location of the divergence is at the interpolation threshold: the number of parameters of the model (elements of Γ) is, to leading order in d, equal to d^2, which matches the number of pretraining samples at $\tau = 1$. Figure 1 confirms this phenomenon in a selection of nonlinear models. We recover a peak in error at the interpolation threshold (given by n), and tracking the location of the interpolation threshold as d increases recovers the quadratic scaling $n \sim d^2$.

Learning transition with increasing pretraining task diversity. It's important to quantify if and when a given model is actually learning in-context, that is, solving a new regression problem by

adapting to the specific structure of the task rather than relying solely on memorized training task vectors. We refer to this phenomenon as *task generalization*.

We posit that a model achieves task generalization when its performance on the ICL test distribution matches its performance on the IDG test distribution. We will thus study the difference between model errors on these two tasks, namely the quantity $g_{\text{task}} = e^{\text{ICL}} - e^{\text{IDG}}$ for a given model or estimator. This difference being large implies the model, performing better on training tasks, has not learned the true task distribution and is not generalising in task. Conversely, a small difference between ICL error and IDG error suggests that the model is leveraging the underlying structure of the task distribution rather than overfitting to and interpolating specific task instances seen in training.

Indeed g_{task} provides a benchmark to determine whether the model's adaptation capabilities exceed memorization, but it's not the whole story. There is a crucial dependence on the parameter $\kappa = k/d$ that controls the diversity of the training task vectors. We will quantify the *rate* at which g_{task} for a given model decreases as κ increases for two inference models: (1) the linear transformer considered thus far, and (2) a memorization prior called the *discrete minimum mean squared error (dMMSE) estimator* as in [18].

Linear transformer. Consider the linear transformer (LT) model $\hat{y} = \langle \Gamma^*, H_Z \rangle$ with ICL and IDG errors given by result 1 and result 2 respectively. How quickly does $g_{\text{task}}^{\text{LT}}$ limit to 0 as $\kappa \to \infty$? By expanding $e^{\text{ICL}} - e^{\text{IDG}}$ in κ we have $g_{\text{task}}^{\text{LT}} = \mathcal{O}\left(\kappa^{-1}\right)$.

dMMSE estimator. We consider the performance of the following estimator, as considered in [18]:

$$w^{\text{dMMSE}} := \sum_{j=1}^{k} w_j e^{-\frac{1}{2\rho}\sum_{i=1}^{\ell}\left(y_i - w_j^\top x_i\right)^2} \Bigg/ \sum_{j=1}^{k} e^{-\frac{1}{2\rho}\sum_{i=1}^{\ell}\left(y_i - w_j^\top x_i\right)^2} . \tag{14}$$

The form of w^{dMMSE} only depends on the training set tasks and so can be called a 'perfect memorizer' or a 'memorization prior.' In Section SI-14 we argue, for large κ, $g_{\text{task}}^{\text{dMMSE}} = \mathcal{O}(\kappa^{-2/d})$.

We conclude that the linear transformer model is a *markedly more efficient task generalizer* than the memorization-prior w^{dMMSE}. The $1/\kappa$ decay in $g_{\text{task}}^{\text{LT}}$ vs the $\approx 1/\kappa^{2/d}$ decay of $g_{\text{task}}^{\text{dMMSE}}$ suggests that the linear transformer *quickly* learns an inference algorithm that generalizes in-context rather than interpolates between training tasks. Simulations are shown in Figure 2a for our linear transformer with theory lines for comparison, and in Figure 2b for a nonlinear transformer model. In both cases we recover the prediction that the transformer architectures are implementing an internal algorithm that can generalise in task much faster than a memorisation prior over the training tasks.

5 Conclusions

In this work, we compute sharp asymptotics for the in-context learning (ICL) performance in a simplified model of ICL for linear regression using linear attention. This exactly solvable model demonstrates a transition in the generalizing capability of the model as the diversity of pretraining tasks increases, echoing empirical findings in full Transformers [18]. Additionally, we observe a sample-wise double descent as the amount of pretraining data increases. Our numerical experiments show that full, nonlinear Transformers exhibit similar behavior in the scaling regime relevant to our solvable model. Our work represents a first step towards a detailed theoretical understanding of the conditions required for ICL to emerge [19].

Finally, our results have some bearing on the broad question of what architectural features are required for ICL [7, 11, 19]. Our work shows that a full Transformer—or indeed even full linear attention—is not required for ICL of linear regression. However, our simplified model retains the structured quadratic pairwise interaction between inputs that is at the heart of the attention mechanism. It is this quadratic interaction that allows the model to solve the ICL regression task, which it does essentially by reversing the data correlation. One would therefore hypothesize that our model is minimal in the sense that further simplifications within this model class would impair its ability to solve this ICL task. In the specific context of regression with isotropic data, a simple point of comparison would be to fix $\Gamma = I_d$, which gives a pretraining-free model that should perform well when the context length is very long. However, this further-reduced model would perform poorly if the covariates of the in-context task are anisotropic. More generally, it would be interesting to investigate when models lacking this precisely-engineered quadratic interaction can learn linear regression in-context, and if they are less sample-efficient than the attention-based models considered here.

Acknowledgements

We thank William L. Tong for helpful discussions regarding numerics. YML was supported by NSF Award CCF-1910410, by the Harvard FAS Dean's Fund for Promising Scholarship, and by a Harvard College Professorship. JAZV and CP were supported by NSF Award DMS-2134157 and NSF CAREER Award IIS-2239780. JAZV is presently supported by a Junior Fellowship from the Harvard Society of Fellows. CP is further supported by a Sloan Research Fellowship. AM acknowledges support from Perimeter Institute, which is supported in part by the Government of Canada through the Department of Innovation, Science and Economic Development and by the Province of Ontario through the Ministry of Colleges and Universities. This work has been made possible in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence. This research was supported in part by grants NSF PHY-1748958 and PHY-2309135 to the Kavli Institute for Theoretical Physics (KITP), through the authors' participation in the Fall 2023 program "Deep Learning from the Perspective of Physics and Neuroscience."

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, 2021.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[5] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training, 2018.

[6] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[8] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[9] Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, et al. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1747–1764, 2022.

[10] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

[11] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=yzkSU5zdwD.

[12] Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 35151–35174. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

[13] Ruiqi Zhang, Jingfeng Wu, and Peter L. Bartlett. In-context learning of a linear transformer block: Benefits of the mlp component and one-step gd initialization, 2024.

[14] Pritam Chandra, Tanmay Kumar Sinha, Kabir Ahuja, Ankit Garg, and Navin Goyal. Towards analyzing self-attention via linear neural network, 2024. URL https://openreview.net/forum?id=4fVuBf5HE9.

[15] Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, and Peter Bartlett. How many pretraining tasks are needed for in-context learning of linear regression? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vSh5ePa0ph.

[16] Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024. URL http://jmlr.org/papers/v25/23-1042.html.

[17] Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning, 2023.

[18] Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 14228–14246. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/2e10b2c2e1aa4f8083c37dfe269873f8-Paper-Conference.pdf.

[19] Gautam Reddy. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=aN4Jf6Cx69.

[20] Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity, 2020.

[21] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3531–3539, 2021.

[22] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR, 2020.

[23] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116 (32):15849–15854, 2019.

[24] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949 – 986, 2022. doi: 10.1214/21-AOS2133. URL https://doi.org/10.1214/21-AOS2133.

[25] Alexander B Atanasov, Jacob A Zavatone-Veth, and Cengiz Pehlevan. Scaling and renormalization in high-dimensional regression. *arXiv preprint arXiv:2405.00592*, 2024.

[26] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015.

[27] László Erdős, Antti Knowles, Horng-Tzer Yau, and Jun Yin. The local semicircle law for a general class of random matrices. *Electronic Journal of Probability*, 18(none):1 – 58, 2013. doi: 10.1214/EJP.v18-2473. URL https://doi.org/10.1214/EJP.v18-2473.

[28] László Erdős and Horng-Tzer Yau. *A dynamical approach to random matrix theory*, volume 28. American Mathematical Soc., 2017.

[29] Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

Supplementary Information

SI-6 Parameter Reduction

Recall that we can express the output of the linear attention mechanism (with full K, Q, V parameters) as

$$\hat{y} = \frac{1}{\ell}\Big\langle x_{\ell+1}, v_{22} M_{11}^\top \sum_{i=1}^{\ell} y_i x_i + v_{22} m_{21} \sum_{i=1}^{\ell} y_i^2 + M_{11}^\top \sum_{i=1}^{\ell+1} x_i x_i^\top v_{21} + m_{21} \sum_{i=1}^{\ell} y_i x_i^\top v_{21} \Big\rangle, \quad (15)$$

where $\langle \cdot, \cdot \rangle$ stands for the standard inner product. We previously argued that the term

$$\frac{1}{\ell} v_{22} M_{11}^\top \sum_{i=1}^{\ell} y_i x_i \quad (16)$$

makes a good prediction for the output. Further, the third term does not depend on outputs y, and thus does not directly contribute to the ICL task that relies on the relationship between x and y. Finally, the last term only considers a one dimensional projection of x onto v_{21}. Because the task vectors w and x are isotropic in the statistical models that we consider, there are no special directions in the problem. Consequently, we expect the optimal v_{21} to be approximately zero by symmetry considerations.

We note that Zhang et al. [16] provide an analysis of population risk (whereas we focus on empirical risk) for a related reduced model in which they set $v_{21} = 0$ and $m_{21} = 0$. Consequently, the predictors they study differ from ours (5) by an additive term. They justify this choice through an optimization argument: if these parameters are initialized to zero, they remain zero under gradient descent optimization of the population risk, given certain conditions.

SI-7 Experimental Details

Our experiments[2] are done with a standard Transformer architecture, where each sample context initially takes the form given by (1). The fully-parameterised linear transformer and softmax-only transformer (which appear in fig. 1c) do not use MLPs. If MLPs are used (e.g. fig. 1b and fig. 1c), the architecture consists of blocks with: (1) a single-head softmax self-attention with $K, Q, V \in \mathbb{R}^{d+1 \times d+1}$ matrices, followed by (2) a two-layer dense MLP with GELU activation and hidden layer of size $d + 1$ [1]. Residual connections are used between the input tokens (padded from dimension d to $d + 1$), the pre-MLP output, and the MLP output. We use a variable number of attention+MLP blocks before returning the final logit corresponding to the $(d + 1, \ell + 1)$th element in the original embedding structure given by (1). The loss function is the mean squared error (MSE) between the predicted label (the output of the model for a given sample Z) and the true value $y_{\ell+1}$. We train the model in an offline setting with n total samples $Z_1, \cdots, Z_n$, divided into 10 batches, using the Adam optimizer [26] with a learning rate 10^{-4} until the training error converges, typically requiring 10000 epochs[3]. The structure of the pretraining and test distributions exactly follows the setup for the ICL task described in Section 2.

SI-8 Figure Details

Figure 1 Figure 1a: Simulated errors are calculated by evaluating the corresponding test error on the corresponding optimised Γ^*. *Parameters:* $d = 100$, $\rho = 0.01$ $\kappa = 0.5$ Averages and standard deviations are computed over 10 runs.
Figure 1b, Figure 1c: Interpolation thresholds shown in fig. 1c were computed empirically by searching for location in τ of sharp increase in value and variance of training error at a fixed number

[2] Code to reproduce all experiments available on github.

[3] Note that larger d models are often trained for less epochs than smaller d models due to early stopping; that said, whether or not early stopping is used in training does not affect either the alignment of error curves in d-scaling nor the qualitative behaviour (double descent in τ and transition in κ).

of gradient steps. The log-log plot demostrating quadratic scaling of n in d was best-fit on the data points plotted. Explicitly, the exponents of d are $a_{\text{full linear}} = 1.87$, $a_{\text{softmax}} = 1.66$, $a_{2\text{ blocks}} = 2.13$, $a_{3\text{ blocks}} = 2.08$. Theory predicts $a = 2$.

Parameters: $\alpha = 1, \kappa = \infty, \rho = 0.01$. For 1b variance shown comes from model trained over different samples of pretraining data; lines show averages over 10 runs and shaded region shows standard deviation.

Figure 3 *Parameters for fig. 2a:* $d = 100$, $\rho = 0.01$. Simulations deviate from theory curve at low κ due to finite size effects. Averages and standard deviations for linear model are computed over 100 runs; dMMSE error is computed numerically over 1000-5000 runs.

Parameters for fig. 2b: $\tau = 10, \alpha = 1, \rho = 0.01$. Variance shown comes from 10 models trained over different samples of pretraining data.

SI-9 Notation

Our derivations will frequently use the vectorization operation, denoted by $\text{vec}(\cdot)$. Note that we shall adopt the *row-major* convention, and thus the rows of A are stacked together to form $\text{vec}(A)$. We also recall the standard identity:

$$\text{vec}(E_1 E_2 E_3) = (E_1 \otimes E_3^\top)\,\text{vec}(E_2), \tag{17}$$

where $\otimes$ denotes the matrix Kronecker product, and E_1, E_2, E_3 are matrices whose dimensions are compatible for the multiplication operation. For any square matrix $A \in \mathbb{R}^{(L+1)\times(L+1)}$, we introduce the notation

$$[M]_{\setminus 0} \in \mathbb{R}^{L\times L} \tag{18}$$

to denote the principal minor of M after removing its first row and column.

Stochastic order notation: In our analysis, we use a concept of high-probability bounds known as *stochastic domination*. This notion, first introduced in [27, 28], provides a convenient way to account for low-probability exceptional events where some bounds may not hold.

We also use the notation $X \simeq Y$ to indicate that two families of random variables X, Y are asymptotically equivalent. Precisely, $X \simeq Y$, if there exists $\varepsilon > 0$ such that for every $D > 0$ we have

$$\mathbb{P}\left[|X - Y| > d^{-\varepsilon}\right] \le d^{-D} \tag{19}$$

for all sufficiently large $d > d_0(\varepsilon, D)$.

SI-10 Moment Calculations and Generalization Errors

For a given set of parameters Γ, its generalization error is defined as

$$e(\Gamma) = \mathbb{E}_{\mathcal{P}_{\text{test}}}\left[\left(y_{\ell+1} - \langle \Gamma, H_Z\rangle\right)^2\right], \tag{20}$$

where $(Z, y_{\ell+1}) \sim \mathcal{P}_{\text{test}}$ is a new sample drawn from the distribution of the test data set. Recall that Z is the input embedding matrix defined in (1) in the main text, and $y_{\ell+1}$ denotes the missing value to be predicted. The goal of this section is to derive an expression for the generalization error $e(\Gamma)$.

Note that the test distribution $\mathcal{P}_{\text{test}}$ crucially depends on the probability distribution of the task vector w used in the linear model. Recall our discussions about Evaluation in 2. For the ICL test task, and for the purposes of analytical tractability, we take $w \sim \text{Unif}(\mathcal{S}^{d-1}(\sqrt{d}))$. In high dimensions, the characterization of ICL error using $w \sim \text{Unif}(\mathcal{S}^{d-1}(\sqrt{d}))$ will be identical to using the original $w \sim \mathcal{N}(0, \mathbb{I})$ used in the main paper body. For the ICL test task, we thus have $w \sim \text{Unif}(\mathcal{S}^{d-1}(\sqrt{d}))$, the uniform distribution on the sphere . In what follows, we slightly abuse the notation by writing $w \sim \mathcal{P}_{\text{test}}$ to indicate that w is sampled from the task vector distribution associated with $\mathcal{P}_{\text{test}}$.

Let w be the task vector used in the input matrix Z. Throughout the paper, we use $\mathbb{E}_w[\cdot]$ to denote the conditional expectation with respect to the randomness in the data vectors $\{x_i\}_{i\in[\ell+1]}$ and the noise $\{\epsilon_i\}_{i\in[\ell+1]}$, with the task vector w kept fixed. We have the following expressions for the first two *conditional* moments of $(H_Z, y_{\ell+1})$.

Lemma 1 (Conditional moments)**.** *Let the task vector $w \in$ be fixed. We have*

$$\mathbb{E}_w\left[y_{\ell+1}\right] = 0, \quad \textit{and} \quad \mathbb{E}_w\left[H_Z\right] = 0. \tag{21}$$

Moreover,

$$\mathbb{E}_w\left[y_{\ell+1} H_Z\right] = \frac{1}{d} w \begin{bmatrix} w^\top, & 1+\rho \end{bmatrix} \tag{22}$$

and

$$\mathbb{E}_w\left[\operatorname{vec}(H_Z)\operatorname{vec}(H_Z)^\top\right] = \frac{(1+\rho)}{d} I_d \otimes \begin{bmatrix} \frac{d}{\ell} I_d + (1+\ell^{-1})(1+\rho)^{-1} w w^\top & (1+2\ell^{-1}) w \\ (1+2\ell^{-1}) w^\top & (1+2\ell^{-1})(1+\rho) \end{bmatrix}. \tag{23}$$

Proof. Using the equivalent representations in (169) and (170), it is straightforward to verify the estimates of the first (conditional) moments in (21). To show (22), we note that

$$H_Z = (d/\ell) z_a z_b^\top, \tag{24}$$

where

$$z_a = M_w \begin{bmatrix} s \\ u \end{bmatrix} \qquad \text{and} \qquad z_b = \begin{bmatrix} M_w h \\ (\theta_w a/\sqrt{d} + \theta_\epsilon)^2/\sqrt{d} + \theta_q^2/\sqrt{d} \end{bmatrix}. \tag{25}$$

Using the representation in (170), we have

$$\mathbb{E}_w\left[y_{\ell+1} H_Z\right] = (d/\ell)\mathbb{E}_w\left[y_{\ell+1} z_a\right] \mathbb{E}_w\left[z_b^\top\right]. \tag{26}$$

Computing the expectations $\mathbb{E}_w\left[y_{\ell+1} z_a\right]$ and $\mathbb{E}_w\left[z_b^\top\right]$ then gives us (22). Next, we show (23). Since z_a and z_b are *independent*,

$$\mathbb{E}\left[\operatorname{vec}(H_Z)\operatorname{vec}(H_Z)^\top\right] = (d/\ell)^2\, \mathbb{E}\left[z_a z_a^\top\right] \otimes \mathbb{E}\left[z_b z_b^\top\right]. \tag{27}$$

The first expectation on the right-hand side is easy to compute. Since M_w is an orthonormal matrix,

$$\mathbb{E}_w\left[z_a z_a^\top\right] = I_d \tag{28}$$

To obtain the second expectation on the right-hand side of the above expression, we can first verify that

$$\mathbb{E}_w\left[M_w h h^\top M_w\right] = \frac{\ell}{d^2}\left[(1+\rho) I_d + \frac{(\ell+1)}{d} w w^\top\right]. \tag{29}$$

Moreover,

$$\mathbb{E}_w\left[M_w h \left((a/\sqrt{d} + \theta_\epsilon)^2/\sqrt{d} + \theta_q^2/\sqrt{d}\right)\right] = \frac{\ell(\ell+2)(1+\rho)}{d^3} w \tag{30}$$

and

$$\mathbb{E}_w\left[\left((a/\sqrt{d} + \theta_\epsilon)^2/\sqrt{d} + \theta_q^2/\sqrt{d}\right)^2\right] = \frac{\ell(\ell+2)(1+\rho)^2}{d^3}. \tag{31}$$

Combining (29), (30), and (31), we have

$$\mathbb{E}\left[z_b z_b^\top\right] = \frac{(\ell/d)^2(1+\rho)}{d} \begin{bmatrix} \frac{d}{\ell} I_d + (1+\ell^{-1})(1+\rho)^{-1} w w^\top & (1+2\ell^{-1}) w \\ (1+2\ell^{-1}) w^\top & (1+2\ell^{-1})(1+\rho) \end{bmatrix}. \tag{32}$$

Substituting (28) and (32) into (23), we reach the formula in (23). □

Proposition 1 (Generalization error)**.** *For a given weight matrix Γ, the generalization error of the linear transformer is*

$$\begin{aligned} e(\Gamma) = \frac{1+\rho}{d} \operatorname{tr}\left(\Gamma \begin{bmatrix} \frac{d}{\ell} I_d + (1+\ell^{-1})(1+\rho)^{-1} R_{\text{test}} & (1+2\ell^{-1}) b_{\text{test}} \\ (1+2\ell^{-1}) b_{\text{test}}^\top & (1+2\ell^{-1})(1+\rho) \end{bmatrix} \Gamma^\top\right) \\ - \frac{2}{d} \operatorname{tr}\left(\Gamma \begin{bmatrix} R_{\text{test}} \\ (1+\rho) b_{\text{test}}^\top \end{bmatrix}\right) + 1 + \rho, \end{aligned} \tag{33}$$

where

$$b_{\text{test}} \coloneqq \mathbb{E}_{w\sim\mathcal{P}_{\text{test}}}[w] \quad \textit{and} \quad R_{\text{test}} \coloneqq \mathbb{E}_{w\sim\mathcal{P}_{\text{test}}}\left[w w^\top\right]. \tag{34}$$

Remark 1. *We use $w \sim \mathcal{P}_{\text{test}}$ to indicate that w is sampled from the task vector distribution associated with $\mathcal{P}_{\text{test}}$. It is then straightforward to check that we want*

$$\textit{(ICL)}: \qquad b_{\text{test}} = 0 \quad \textit{and} \quad R_{\text{test}} = I_d. \tag{35}$$

Proof. Recall the definition of the generalization error in (20). We start by writing

$$e(\Gamma) = \operatorname{vec}(\Gamma)^\top \mathbb{E}\left[\operatorname{vec}(H_Z)\operatorname{vec}(H_Z)^\top\right]\operatorname{vec}(\Gamma) - 2\operatorname{vec}(\Gamma)^\top \operatorname{vec}(\mathbb{E}\left[y_{N+1}H_Z\right]) + \mathbb{E}\left[y_{\ell+1}^2\right], \tag{36}$$

where H_Z is a matrix in the form of (6) and H_Z is independent of Γ. Since $y_{\ell+1} = x_{\ell+1}^\top w + \epsilon$, with $\epsilon \sim \mathcal{N}(0,\rho)$ denoting the noise, it is straightforward to check that

$$\mathbb{E}\left[y_{\ell+1}^2\right] = 1 + \rho. \tag{37}$$

Using the moment estimate (23) in Lemma 1 and the identity (17), we have

$$\begin{aligned}&\operatorname{vec}(\Gamma)^\top \mathbb{E}\left[\operatorname{vec}(H_Z)\operatorname{vec}(H_Z)^\top\right]\operatorname{vec}(\Gamma)\\ &\quad= \frac{1+\rho}{d}\operatorname{tr}\left(\Gamma\begin{bmatrix}\frac{d}{\ell}I_d + (1+\ell^{-1})(1+\rho)^{-1}R_{\text{test}} & (1+2\ell^{-1})b_{\text{test}}\\ (1+2\ell^{-1})b_{\text{test}}^\top & (1+2\ell^{-1})(1+\rho)\end{bmatrix}\Gamma^\top\right).\end{aligned} \tag{38}$$

Moreover, by (22),

$$\operatorname{vec}(\Gamma)^\top \operatorname{vec}\left(\mathbb{E}\left[y_{\ell+1}H_Z\right]\right) = \frac{1}{d}\operatorname{tr}\left(\Gamma\begin{bmatrix}R_{\text{test}}\\ (1+\rho)b_{\text{test}}^\top\end{bmatrix}\right). \tag{39}$$

□

Corollary 1. *For a given set of parameters Γ, its generalization error can be written as*

$$e(\Gamma) = \frac{1}{d}\operatorname{tr}\left(\Gamma B_{\text{test}}\Gamma^\top\right) - \frac{2}{d}\operatorname{tr}\left(\Gamma A_{\text{test}}^\top\right) + (1+\rho) + \mathcal{E}, \tag{40}$$

where

$$A_{\text{test}} \coloneqq \begin{bmatrix}R_{\text{test}} & (1+\rho)b_{\text{test}}\end{bmatrix}, \tag{41}$$

$$B_{\text{test}} \coloneqq \begin{bmatrix}\frac{1}{\alpha}(1+\rho)I_d + R_{\text{test}} & (1+\rho)b_{\text{test}}\\ (1+\rho)b_{\text{test}}^\top & (1+\rho)^2\end{bmatrix}, \tag{42}$$

and $R_{\text{test}}, b_{\text{test}}$ are as defined in (34). Moreover, $\mathcal{E}$ denotes an "error" term such that

$$|\mathcal{E}| \le \frac{C_{\alpha,\rho}\max\left\{\|R_{\text{test}}\|_{\mathsf{op}}, \|b_{\text{test}}\|, 1\right\}\left(\|\Gamma\|_{\mathsf{F}}^2/d\right)}{d}, \tag{43}$$

where $C_{\alpha,\rho}$ is some constant that only depends on α and ρ.

Proof. Let

$$\Delta = \begin{bmatrix}\frac{d}{\ell}(1+\rho)I_d + (1+\ell^{-1})R_{\text{test}} & (1+2\ell^{-1})(1+\rho)b_{\text{test}}\\ (1+2\ell^{-1})(1+\rho)b_{\text{test}}^\top & (1+2\ell^{-1})(1+\rho)^2\end{bmatrix} - B_{\text{test}}. \tag{44}$$

It is straightforward to check that

$$\mathcal{E} = \frac{1}{d}\operatorname{tr}\left(\Gamma\Delta\Gamma^\top\right) \tag{45}$$

$$= \frac{1}{d}\operatorname{vec}(\Gamma)^\top(I_d \otimes \Delta)\operatorname{vec}(\Gamma) \tag{46}$$

$$\le \|\Delta\|_{\mathsf{op}}\frac{\|\Gamma\|_{\mathsf{F}}^2}{d}. \tag{47}$$

The bound in (43) follows from the estimate that $\|\Delta\|_{\mathsf{op}} \le C_{\alpha,\rho}\max\left\{\|R_{\text{test}}\|_{\mathsf{op}}, \|b_{\text{test}}\|, 1\right\}/d$. □

Remark 2. *Consider the optimal weight matrix Γ^* obtained by solving the ridge regression problem in (7). Since Γ^* is the optimal solution of (7), we must have*

$$\frac{n}{d}\lambda\|\Gamma^*\|_{\mathsf{F}}^2 \le \sum_{\mu\in[n]}(y_{\ell+1}^{\mu})^2, \tag{48}$$

where the right-hand side is the value of the objective function of (7) when we choose Γ to be the all-zero matrix. It follows that

$$\frac{\|\Gamma^*\|_{\mathsf{F}}^2}{d} \le \frac{\sum_{\mu\in[n]}(y_{\ell+1}^{\mu})^2}{\lambda\, n}. \tag{49}$$

By the law of large numbers, $\frac{\sum_{\mu\in[n]} y_\mu^2}{n} \to 1+\rho$ as $n\to\infty$. Thus, $\|\Gamma^*\|_{\mathsf{F}}^2/d$ is asymptotically bounded by the constant $(1+\rho)/\lambda$. Furthermore, it is easy to check that $\|R_{\rm test}\|_{\mathsf{op}} = \mathcal{O}(1)$ and $\|b_{\rm test}\| = \mathcal{O}(1)$ for the ICL task [see (35)]. It then follows from Corollary 1 that the generalization error associated with the optimal parameters Γ^* is asymptotically determined by the first three terms on the right-hand side of (40).

SI-11 Analysis of Ridge Regression: Extended Resolvent Matrices

We see from Corollary 1 and Remark 2 that the two key quantities in determining the generalization error $e(\Gamma^*)$ are

$$\frac{1}{d}\operatorname{tr}(\Gamma^* A_{\rm test}^\top) \qquad \text{and} \qquad \frac{1}{d}\operatorname{tr}(\Gamma^* B_{\rm test}(\Gamma^*)^\top), \tag{50}$$

where $A_{\rm test}$ and $B_{\rm test}$ are the matrices defined in (41) and (42), respectively. In this section, we show that the two quantities in (50) can be obtained by studying a parameterized family of extended resolvent matrices.

To start, we observe that the ridge regression problem in (5) admits the following closed-form solution:

$$\operatorname{vec}(\Gamma^*) = G\left(\textstyle\sum_{\mu\in[n]} y_\mu \operatorname{vec}(H_\mu)\right)/d, \tag{51}$$

where G is a resolvent matrix defined as

$$G = \left(\textstyle\sum_{\mu\in[n]} \operatorname{vec}(H_\mu)\operatorname{vec}(H_\mu)^\top/d + \tau\lambda I\right)^{-1}. \tag{52}$$

For our later analysis of the generalization error, we need to consider a more general, "parameterized" version of G, defined as

$$G(\pi) = \left(\textstyle\sum_{\mu\in[n]} \operatorname{vec}(H_\mu)\operatorname{vec}(H_\mu)^\top/d + \pi\Pi + \tau\lambda I\right)^{-1}, \tag{53}$$

where $\Pi \in \mathbb{R}^{(d^2+d)\times(d^2+d)}$ is a symmetric positive-semidefinite matrix and π is a nonnegative scalar. The original resolvent G in (52) is a special case, corresponding to $\pi = 0$.

The objects in (51) and (53) are the submatrices of an *extended* resolvent matrix, which we construct as follows. For each $\mu\in[n]$, let

$$z_\mu = \begin{bmatrix} y_\mu/d \\ \operatorname{vec}(H_\mu)/\sqrt{d}\end{bmatrix} \tag{54}$$

be an (d^2+d+1)-dimensional vector. Let

$$\Pi_e = \begin{bmatrix} 0 & \\ & \Pi\end{bmatrix}, \tag{55}$$

where Π is the $(d^2+d)\times(d^2+d)$ matrix in (53). Define an extended resolvent matrix

$$G_e(\pi) = \frac{1}{\sum_{\mu\in[n]} z_\mu z_\mu^\top + \pi\Pi_e + \tau\lambda I}. \tag{56}$$

By block-matrix inversion, it is straightforward to check that

$$G_e(\pi) = \begin{bmatrix} c(\pi) & -c(\pi)q^\top(\pi) \\ -c(\pi)q(\pi) & G(\pi) + c(\pi)q(\pi)q^\top(\pi)\end{bmatrix}, \tag{57}$$

where

$$q(\pi) \coloneqq \frac{1}{d^{3/2}} G(\pi) \left(\textstyle\sum_{\mu \in [n]} y_\mu \operatorname{vec}(H_\mu) \right) \tag{58}$$

is a vector in $\mathbb{R}^{d(d+1)}$, and $c(\pi)$ is a scalar such that

$$\frac{1}{c(\pi)} = \frac{1}{d^2} \sum_{\mu \in [n]} y_\mu^2 + \tau\lambda - \frac{1}{d^3} \sum_{\mu,\nu \in [n]} y_\mu y_\nu \operatorname{vec}(H_\mu)^\top G(\pi) \operatorname{vec}(H_\nu). \tag{59}$$

By comparing (58) with (51), we see that

$$\operatorname{vec}(\Gamma^*) = \sqrt{d}\, q(0). \tag{60}$$

Moreover, as shown in the following lemma, the two key quantities in (50) can also be obtained from the extended resolvent $G_e(\pi)$.

Lemma 2. *For any matrix $A \in \mathbb{R}^{d \times (d+1)}$,*

$$\frac{1}{d} \operatorname{tr}(\Gamma^* A^\top) = \frac{-1}{c(0)\sqrt{d}} \begin{bmatrix} 0 & \operatorname{vec}(A)^T \end{bmatrix} G_e(0) e_1, \tag{61}$$

where e_1 denotes the first natural basis vector in $\mathbb{R}^{d^2+d+1}$. Moreover, for any symmetric and positive semidefinite matrix $B \in \mathbb{R}^{(d+1)\times(d+1)}$, if we set

$$\Pi = I_d \otimes B \tag{62}$$

in (55), then

$$\frac{1}{d} \operatorname{tr}(\Gamma^* B (\Gamma^*)^\top) = \frac{\mathrm{d}}{\mathrm{d}\pi} \left(\frac{1}{c(\pi)} \right) \Big|_{\pi=0}. \tag{63}$$

Proof. The identity (61) follows immediately from the block form of $G_e(\pi)$ in (57) and the observation in (60). To show (63), we take the derivative of $1/c(\pi)$ with respect to π. From (59), and using the identity

$$\frac{\mathrm{d}}{\mathrm{d}\pi} G(\pi) = -G(\pi) \Pi G(\pi), \tag{64}$$

we have

$$\frac{\mathrm{d}}{\mathrm{d}\pi} \left(\frac{1}{c(\pi)} \right) = \frac{1}{d^3} \sum_{\mu,\nu \in [n]} y_\mu y_\nu \operatorname{vec}(H_\mu)^\top G(\pi) \Pi G(\pi) \operatorname{vec}(H_\nu) \tag{65}$$

$$= q^\top(\pi) \Pi q(\pi). \tag{66}$$

Thus, by (60),

$$\frac{\mathrm{d}}{\mathrm{d}\pi} \left(\frac{1}{c(\pi)} \right) \Big|_{\pi=0} = \frac{1}{d} \left(\operatorname{vec}(\Gamma^*) \right)^\top \Pi \operatorname{vec}(\Gamma^*) \tag{67}$$

$$= \frac{1}{d} \left(\operatorname{vec}(\Gamma^*) \right)^\top (I_d \otimes B) \operatorname{vec}(\Gamma^*). \tag{68}$$

Applying the identity in (17) to the right-hand side of the above equation, we reach (63). □

Remark 3. *To lighten the notation, we will often write $G_e(\pi)$ [resp. $G(\pi)$] as G_e [resp. G], leaving their dependence on the parameter π implicit.*

Remark 4. *In light of (62) and (63), we will always choose*

$$\Pi = I_d \otimes B_{\text{test}}, \tag{69}$$

where B_{test} is the matrix defined in (42).

SI-12 An Asymptotic Equivalent of the Extended Resolvent Matrix

In this section, we derive an asymptotic equivalent of the extended resolvent G_e defined in (56). From this equivalent version, we can then obtain the asymptotic limits of the right-hand sides of (61) and (63). Our analysis relies on non-rigorous but technically sound heuristic arguments from random matrix theory. Therefore, we refer to our theoretical predictions as *results* rather than propositions.

Recall that there are k unique task vectors $\{w_i\}_{i\in[k]}$ in the training set. Let

$$b_{\rm tr} \coloneqq \frac{1}{k}\sum_{i\in[k]} w_i \quad \text{and} \quad R_{\rm tr} \coloneqq \frac{1}{k}\sum_{i\in[k]} w_i w_i^\top \tag{70}$$

denote the empirical mean and correlation matrix of these k regression vectors, respectively. Define

$$A_{\rm tr} \coloneqq \begin{bmatrix} R_{\rm tr} & (1+\rho)b_{\rm tr}\end{bmatrix}. \tag{71}$$

and

$$E_{\rm tr} \coloneqq \begin{bmatrix} \frac{(1+\rho)}{\alpha} I_d + R_{\rm tr} & (1+\rho)b_{\rm tr} \\ (1+\rho)b_{\rm tr}^\top & (1+\rho)^2 \end{bmatrix}. \tag{72}$$

Definition 1. *Consider the extended resolvent $G_e(\pi)$ in (56), with Π_e chosen in the forms of (55) and (69). Let $\widetilde{G}_e$ be another matrix of the same size as $G_e(\pi)$. We say that $\widetilde{G}_e$ and $G_e(\pi)$ are asymptotically equivalent, if the following conditions hold.*

(1) For any two deterministic and unit-norm vectors $u, v \in \mathbb{R}^{d^2+d+1}$,

$$u^\top G_e(\pi) v \simeq u^\top \widetilde{G}_e v, \tag{73}$$

where $\simeq$ is the asymptotic equivalent notation defined in (19).

(2) Let $A_{\rm tr} = \begin{bmatrix} R_{\rm tr} & (1+\rho)b_{\rm tr}\end{bmatrix}$. For any deterministic, unit-norm vector $v \in \mathbb{R}^{d^2+d+1}$,

$$\frac{1}{\sqrt{d}}\begin{bmatrix} 0 & \mathrm{vec}(A_{\rm tr})^\top\end{bmatrix} G_e(\pi) v \simeq \frac{1}{\sqrt{d}}\begin{bmatrix} 0 & \mathrm{vec}(A_{\rm tr})^\top\end{bmatrix} \widetilde{G}_e v. \tag{74}$$

(3) Recall the notation introduced in (18). We have

$$\frac{1}{d^2}\,\mathrm{tr}\left(\left[G_e(\pi)\right]_{\setminus 0}\cdot[I\otimes E_{\rm tr}]\right) = \frac{1}{d^2}\,\mathrm{tr}\left(\left[\widetilde{G}_e\right]_{\setminus 0}\cdot[I\otimes E_{\rm tr}]\right) + \mathcal{O}_{\prec}(d^{-1/2}), \tag{75}$$

where $\left[G_e(\pi)\right]_{\setminus 0}$ and $\left[\mathcal{G}_e(\pi)\right]_{\setminus 0}$ denote the principal minors of $G_e(\pi)$ and $\mathcal{G}_e(\pi)$, respectively.

Result 3. *Let χ_π denote the unique positive solution to the equation*

$$\chi_\pi = \frac{1}{d}\,\mathrm{tr}\left[\left(\frac{\tau}{1+\chi_\pi}E_{\rm tr} + \pi B_{\rm test} + \lambda\tau I_d\right)^{-1} E_{\rm tr}\right], \tag{76}$$

where $B_{\rm test}$ is the positive-semidefinite matrix in (42), with $b_{\rm test}, R_{\rm test}$ chosen according to (35). The extended resolvent $G_e(\pi)$ in (56) is asymptotically equivalent to

$$\mathcal{G}_e(\pi) \coloneqq \left(\frac{\tau}{1+\chi_\pi}\begin{bmatrix} 1+\rho & \frac{1}{\sqrt{d}}\mathrm{vec}\left(\begin{bmatrix} R_{\rm tr} & (1+\rho)b_{\rm tr}\end{bmatrix}\right)^\top \\ \frac{1}{\sqrt{d}}\mathrm{vec}\left(\begin{bmatrix} R_{\rm tr} & (1+\rho)b_{\rm tr}\end{bmatrix}\right) & I_d\otimes E_{\rm tr}\end{bmatrix} + \pi\Pi_e + \tau\lambda I\right)^{-1}, \tag{77}$$

in the sense of Definition 1. In the above expression, Π_e is the matrix in (55) with $\Pi = I_d \otimes B_{\rm test}$.

In what follows, we present the steps in reaching the asymptotic equivalent $\mathcal{G}_e(\pi)$ given in (77). To start, let $G_e^{[\mu]}$ to denote a "leave-one-out" version of G_e, defined as

$$G_e^{[\mu]} = \frac{1}{\sum_{\nu\neq\mu} z_\nu z_\nu^\top + \pi\Pi_e + \tau\lambda I}. \tag{78}$$

By (56), we have

$$G_e\left(\sum_{\mu\in[n]} z_\mu z_\mu^\top + \pi\Pi_e + \tau\lambda I\right) = I. \tag{79}$$

Applying the Woodbury matrix identity then gives us

$$\sum_{\mu\in[n]} \frac{1}{1 + z_\mu^\top G_e^{[\mu]} z_\mu} G_e^{[\mu]} z_\mu z_\mu^\top + G_e(\pi\Pi_e + \tau\lambda I) = I. \tag{80}$$

To proceed, we study the quadratic form $z_\mu^\top G_e^{[\mu]} z_\mu$. Let w_μ denotes the task vector associated with z_μ. Conditioned on w_μ and G_e^μ, the quadratic form $z_\mu^\top G_e^{[\mu]} z_\mu$ concentrates around its *conditional expectation* with respect to the remaining randomness in z_μ. Specifically,

$$z_\mu^\top G_e^{[\mu]} z_\mu = \chi^\mu(w_\mu) + \mathcal{O}_\prec(d^{-1/2}), \tag{81}$$

where

$$\chi^\mu(w_\mu) \coloneqq \frac{1}{d^2} \operatorname{tr}\left([G_e^\mu]_{\setminus 0} \cdot [I \otimes E(w_\mu)]\right), \tag{82}$$

and

$$E(w) \coloneqq \begin{bmatrix} \frac{1+\rho}{\alpha} I_d + ww^\top & (1+\rho)w \\ (1+\rho)w^\top & (1+\rho)^2 \end{bmatrix}. \tag{83}$$

Substituting $z_\mu^\top G_e^{[\mu]} z_\mu$ in (80) by $\chi^\mu(w_\mu)$, we get

$$\sum_{\mu\in[n]} \frac{1}{1 + \chi^\mu(w_\mu)} G_e^{[\mu]} z_\mu z_\mu^\top + G_e(\pi\Pi_e + \tau\lambda I) = I + \Delta_1, \tag{84}$$

where

$$\Delta_1 \coloneqq \sum_{\mu\in[n]} \frac{z_\mu^\top G_e^{[\mu]} z_\mu - \chi^\mu(w_\mu)}{(1 + \chi^\mu(w_\mu))(1 + z_\mu^\top G_e^{[\mu]} z_\mu)} G_e^{[\mu]} z_\mu z_\mu^\top \tag{85}$$

is a matrix that captures the approximation error of the above substitution.

Next, we replace $z_\mu z_\mu^\top$ on the left-hand side of (84) by its *conditional* expectation $\mathbb{E}_{w_\mu}\left[z_\mu z_\mu^\top\right]$, conditioned on the task vector w_μ. This allows us to rewrite (84) as

$$\sum_{\mu\in[n]} \frac{1}{1 + \chi^\mu(w_\mu)} G_e^{[\mu]} \mathbb{E}_{w_\mu}\left[z_\mu z_\mu^\top\right] + G_e(\pi\Pi_e + \tau\lambda I) = I + \Delta_1 + \Delta_2, \tag{86}$$

where

$$\Delta_2 \coloneqq \sum_{\mu\in[n]} \frac{1}{1 + \chi^\mu(w_\mu)} G_e^{[\mu]} \left(\mathbb{E}_{w_\mu}\left[z_\mu z_\mu^\top\right] - z_\mu z_\mu^\top\right) \tag{87}$$

captures the corresponding approximation error. Recall the definition of z_μ in (54). Using the moment estimates in Lemma 1, we have

$$\mathbb{E}_{w_\mu}\left[z_\mu z_\mu^\top\right] = \frac{1}{d^2}\begin{bmatrix} 1+\rho & \frac{1}{\sqrt{d}} w_\mu^\top \otimes \begin{bmatrix} w_\mu^\top & 1+\rho \end{bmatrix} \\ \frac{1}{\sqrt{d}} w_\mu \otimes \begin{bmatrix} w_\mu \\ 1+\rho \end{bmatrix} & I_d \otimes E(w_\mu) \end{bmatrix} + \frac{1}{d^2}\begin{bmatrix} 0 & \\ & I_d \otimes \mathcal{E}_\mu \end{bmatrix}, \tag{88}$$

where $E(w_\mu)$ is the matrix defined in (83) and

$$\mathcal{E}_\mu = \frac{1}{\ell}\begin{bmatrix} w_\mu w_\mu^\top & 2(1+\rho)w_\mu \\ 2(1+\rho)w_\mu^\top & 2(1+\rho)^2 \end{bmatrix}. \tag{89}$$

Replacing the conditional expectation $\mathbb{E}_{w_\mu}\left[z_\mu z_\mu^\top\right]$ in (86) by the main (i.e. the first) term on the right-hand side of (88), we can transform (86) to

$$\frac{\tau}{n} \sum_{\mu\in[n]} \frac{1}{1 + \chi^\mu(w_\mu)} G_e^{[\mu]} \begin{bmatrix} 1+\rho & \frac{1}{\sqrt{d}} w_\mu^\top \otimes \begin{bmatrix} w_\mu^\top & 1+\rho \end{bmatrix} \\ \frac{1}{\sqrt{d}} w_\mu \otimes \begin{bmatrix} w_\mu \\ 1+\rho \end{bmatrix} & I_d \otimes E(w_\mu) \end{bmatrix} + G_e(\pi\Pi_e + \tau\lambda I) = I + \Delta_1 + \Delta_2 + \Delta_3, \tag{90}$$

where we recall $\tau = n/d^2$, and we use Δ_3 to capture the approximation error associated with $\mathcal{E}_\mu$.

Next, we replace the "leave-one-out" terms G_e^μ and $\chi^\mu(w_\mu)$ in (90) by their "full" versions. Specifically, we replace G_e^μ by G_e, and $\chi^\mu(w_\mu)$ by

$$\chi(w_\mu) \coloneqq \frac{1}{d^2} \operatorname{tr}\left([G_e]_{\setminus 0} \cdot \big[I \otimes E(w_\mu) \big] \right). \tag{91}$$

It is important to note the difference between (82) and (91): the former uses G_e^μ and the latter G_e. After these replacements and using Δ_4 to capture the approximation errors, we have

$$G_e \left(\frac{\tau}{n} \sum_{\mu \in [n]} \frac{1}{1+\chi(w_\mu)} \begin{bmatrix} 1+\rho & \frac{1}{\sqrt{d}} w_\mu^\top \otimes \begin{bmatrix} w_\mu^\top & 1+\rho \end{bmatrix} \\ \frac{1}{\sqrt{d}} w_\mu \otimes \begin{bmatrix} w_\mu \\ 1+\rho \end{bmatrix} & I_d \otimes E(w_\mu) \end{bmatrix} + \pi \Pi_e + \tau\lambda I \right) = I + \sum_{j \le 4} \Delta_j. \tag{92}$$

Recall that there are k unique task vectors $\{w_i\}_{1 \le i \le k}$ in the training set consisting of n input samples. Each sample is associated with one of these task vectors, sampled uniformly from the set $\{w_i\}_{1 \le i \le k}$. In our analysis, we shall assume that k divides n and that each unique task vector is associated with exactly n/k input samples. (We note that this assumption merely serves to simplify the notation. The asymptotic characterization of the random matrix G_e remains the same even without this assumption.) Observe that there are only k unique terms in the sum on the left-hand side of (92). Thus,

$$G_e \left(\frac{\tau}{k} \sum_{i \in [k]} \frac{1}{1+\chi(w_i)} \begin{bmatrix} 1+\rho & \frac{1}{\sqrt{d}} w_i^\top \otimes \begin{bmatrix} w_i^\top & 1+\rho \end{bmatrix} \\ \frac{1}{\sqrt{d}} w_i \otimes \begin{bmatrix} w_i \\ 1+\rho \end{bmatrix} & I_d \otimes E(w_i) \end{bmatrix} + \pi \Pi_e + \tau\lambda I \right) = I + \sum_{j \le 4} \Delta_j. \tag{93}$$

So far, we have been treating the k task vectors $\{w_i\}_{i \in [k]}$ as fixed vectors, only using the randomness in the input samples that are associated with the data vectors $\{x_i^\mu\}$. To further simplify our asymptotic characterization, we take advantage of the fact that $\{w_i\}_{i \in [k]}$ are independently sampled from $\mathrm{Unif}(\mathcal{S}^{d-1}(\sqrt{d}))$. To that end, we can first show that $\chi(w_i)$ in (91) concentrates around its expectation. Specifically,

$$\chi(w_i) = \mathbb{E}\left[\frac{1}{d^2} \operatorname{tr}\left([G_e]_{\setminus 0} \cdot \big[I \otimes E(w_i) \big] \right) \right] + \mathcal{O}_\prec(d^{-1/2}). \tag{94}$$

By symmetry, we must have

$$\mathbb{E}\left[\frac{1}{d^2} \operatorname{tr}\left([G_e]_{\setminus 0} \cdot \big[I \otimes E(w_i) \big] \right) \right] = \mathbb{E}\left[\frac{1}{d^2} \operatorname{tr}\left([G_e]_{\setminus 0} \cdot \big[I \otimes E(w_j) \big] \right) \right] \tag{95}$$

for any $1 \le i < j \le k$. It follows that $\big|\chi(w_i) - \chi(w_j)\big| = \mathcal{O}_\prec(d^{-1/2})$, and thus, by a union bound,

$$\max_{i \in [k]} \big|\chi(w_{k_1}) - \widehat{\chi}_{\mathrm{ave}}\big| = \mathcal{O}_\prec(d^{-1/2}), \tag{96}$$

where

$$\widehat{\chi}_{\mathrm{ave}} \coloneqq \frac{1}{k} \sum_{i \in [k]} \chi(w_i). \tag{97}$$

Upon substituting (91) into (97), it is straightforward to verify the following characterization of $\widehat{\chi}_{\mathrm{ave}}$:

$$\widehat{\chi}_{\mathrm{ave}} = \frac{1}{d^2} \operatorname{tr}\left([G_e]_{\setminus 0} \cdot [I \otimes E_{\mathrm{tr}}] \right). \tag{98}$$

The estimate in (96) prompts us to replace the terms $\chi(w_i)$ in the right-hand side of (93) by the common value $\widehat{\chi}_{\mathrm{ave}}$. As before, we introduce a matrix Δ_5 to capture the approximation error associated with this step. Using the newly introduced notation E_{tr}, b_{tr} and R_{tr} in (72) and (70), we

can then simplify (93) as

$$G_e\left(\frac{\tau}{1+\widehat{\chi}_{\rm ave}}\begin{bmatrix}1+\rho & \frac{1}{\sqrt{d}}\operatorname{vec}\left(\begin{bmatrix}R_{\rm tr} & (1+\rho)b_{\rm tr}\end{bmatrix}\right)^\top\\ \frac{1}{\sqrt{d}}\operatorname{vec}\left(\begin{bmatrix}R_{\rm tr} & (1+\rho)b_{\rm tr}\end{bmatrix}\right) & I_d\otimes E_{\rm tr}\end{bmatrix}+\pi\Pi_e+\tau\lambda I\right)$$
$$=I+\sum_{1\le j\le 5}\Delta_j. \tag{99}$$

Define

$$\widehat{\mathcal{G}}_e(\pi)\coloneqq\left(\frac{\tau}{1+\widehat{\chi}_{\rm ave}}\begin{bmatrix}1+\rho & \frac{1}{\sqrt{d}}\operatorname{vec}\left(\begin{bmatrix}R_{\rm tr} & (1+\rho)b_{\rm tr}\end{bmatrix}\right)^\top\\ \frac{1}{\sqrt{d}}\operatorname{vec}\left(\begin{bmatrix}R_{\rm tr} & (1+\rho)b_{\rm tr}\end{bmatrix}\right) & I_d\otimes E_{\rm tr}\end{bmatrix}+\pi\Pi_e+\tau\lambda I\right)^{-1}. \tag{100}$$

Then

$$G_e=\widehat{\mathcal{G}}_e(\pi)+\underbrace{\widehat{\mathcal{G}}_e(\pi)\left(\Delta_1+\Delta_2+\Delta_3+\Delta_4+\Delta_5\right)}_{\text{approximation errors}}. \tag{101}$$

Remark 5. *We claim that $\widehat{\mathcal{G}}_e$ is asymptotically equivalent to G_e, in the sense of Definition 1. Given (101), proving this claim requires showing that, for $j=1,2,\dots,5$,*

$$u^\top\left(\widehat{\mathcal{G}}_e(\pi)\Delta_j\right)v\simeq 0, \tag{102a}$$

$$\frac{1}{\sqrt{d}}\begin{bmatrix}0 & \operatorname{vec}(A_{\rm tr})^\top\end{bmatrix}\left(\widehat{\mathcal{G}}_e(\pi)\Delta_j\right)v\simeq 0, \tag{102b}$$

and

$$\frac{1}{d^2}\operatorname{tr}\left(\left[\widehat{\mathcal{G}}_e(\pi)\Delta_j\right]_{\setminus 0}\cdot[I\otimes E_{\rm tr}]\right)\simeq 0, \tag{102c}$$

for any deterministic and unit-norm vectors u,v and for $A_{\rm tr}=\begin{bmatrix}R_{\rm tr} & (1+\rho)b_{\rm tr}\end{bmatrix}$.

We note the equivalent matrix $\widehat{\mathcal{G}}_e(\pi)$ still involves one scalar $\widehat{\chi}_{\rm ave}$ that depends on the original resolvent $G_e(\pi)$. Next, we show that $\widehat{\chi}_{\rm ave}$ can be replaced by χ_π, the unique positive solution to (76). To that end, we recall the characterization in (98). Using the claim that $G_e(\pi)$ and $\widehat{\mathcal{G}}_e(\pi)$ are asymptotically equivalent (in particular, in the sense of (75)), we have

$$\widehat{\chi}_{\rm ave}\simeq\frac{1}{d^2}\operatorname{tr}\left(\left[\widehat{\mathcal{G}}_e(\pi)\right]_{\setminus 0}\cdot[I\otimes E_{\rm tr}]\right). \tag{103}$$

To compute the first term on the right-hand side of the above estimate, we directly invert the block matrix $\widehat{\mathcal{G}}_e(\pi)$ in (100). Recall that Π_e is chosen in the forms of (55) and (62). It is then straightforward to verify that

$$\widehat{\mathcal{G}}_e=\begin{bmatrix}\bar{c} & -\bar{c}\,\bar{q}^\top\\ -\bar{c}\,\bar{q} & I\otimes F_E(\widehat{\chi}_{\rm ave})+\bar{c}\,\bar{q}\bar{q}^\top\end{bmatrix}, \tag{104}$$

where $F_E(\chi)$ is a matrix valued function such that

$$F_E(\chi)=\left(\frac{\tau}{1+\chi}E_{\rm tr}+\pi B+\lambda\tau I_{d+1}\right)^{-1}, \tag{105}$$

$$\bar{q}=\frac{\tau}{(1+\widehat{\chi}_{\rm ave})\sqrt{d}}\operatorname{vec}\left(\begin{bmatrix}R_{\rm tr} & (1+\rho)b_{\rm tr}\end{bmatrix}F_E(\widehat{\chi}_{\rm ave})\right), \tag{106}$$

and

$$1/\bar{c}=\frac{\tau(1+\rho)}{1+\widehat{\chi}_{\rm ave}}+\lambda\tau-\frac{\tau^2}{(1+\widehat{\chi}_{\rm ave})^2 d}\operatorname{tr}\left(\begin{bmatrix}R_{\rm tr} & (1+\rho)b_{\rm tr}\end{bmatrix}F_E(\widehat{\chi}_{\rm ave})\begin{bmatrix}R_{\rm tr} & (1+\rho)b_{\rm tr}\end{bmatrix}^\top\right). \tag{107}$$

Using (104), we can now write the equation (103) as

$$\widehat{\chi}_{\rm ave} \simeq \frac{1}{d}\,\mathrm{tr}\left(F_E(\widehat{\chi}_{\rm ave})E_{\rm tr}\right) + \frac{\bar{c}\,\tau^2}{(1+\widehat{\chi}_{\rm ave})^2 d^3}\,\mathrm{tr}\left(\begin{bmatrix}R_{\rm tr} & (1+\rho)b_{\rm tr}\end{bmatrix} F_E(\widehat{\chi}_{\rm ave})E_{\rm tr}F_E(\widehat{\chi}_{\rm ave}) \begin{bmatrix}R_{\rm tr} & (1+\rho)b_{\rm tr}\end{bmatrix}^\top\right). \tag{108}$$

The second term on the right-hand side of (108) is negligible. Indeed,

$$\begin{aligned}&\mathrm{tr}\left(\begin{bmatrix}R_{\rm tr} & (1+\rho)b_{\rm tr}\end{bmatrix} F_E(\widehat{\chi}_{\rm ave})E_{\rm tr}F_E(\widehat{\chi}_{\rm ave}) \begin{bmatrix}R_{\rm tr} & (1+\rho)b_{\rm tr}\end{bmatrix}^\top\right)\\ &\qquad\le \left\|F_E(\widehat{\chi}_{\rm ave})E_{\rm tr}F_E(\widehat{\chi}_{\rm ave})\right\|_{\mathsf{op}}\left(\|R_{\rm tr}\|_{\mathsf{F}}^2 + (1+\rho)^2\|b_{\rm tr}\|^2\right).\end{aligned} \tag{109}$$

By construction, $\left\|F_E(\widehat{\chi}_{\rm ave})\right\|_{\mathsf{op}} \le (\lambda\tau)^{-1}$. Moreover, since the task vectors $\{w_i\}_{i\in[k]}$ are independent vectors sampled from $\mathrm{Unif}(\mathcal{S}^{d-1}(\sqrt{d}))$, it is easy to verify that

$$\|E_{\rm tr}\|_{\mathsf{op}} = \mathcal{O}_{\prec}(1), \qquad \|R_{\rm tr}\|_{\mathsf{F}} = \mathcal{O}_{\prec}(\sqrt{d}) \qquad \text{and} \qquad \|b_{\rm tr}\|_2 = \mathcal{O}_{\prec}(1). \tag{110}$$

Finally, since $\bar{c}$ is an element of $\widehat{\mathcal{G}}_e$, we must have $|\bar{c}| \le \left\|\widehat{\mathcal{G}}_e\right\|_{\mathsf{op}} \le (\tau\lambda)^{-1}$. Combining these estimates gives us

$$\frac{\bar{c}\,\tau^2}{(1+\widehat{\chi}_{\rm ave})^2 d^3}\,\mathrm{tr}\left(\begin{bmatrix}R_{\rm tr} & (1+\rho)b_{\rm tr}\end{bmatrix} F_E(\widehat{\chi}_{\rm ave})E_{\rm tr}F_E(\widehat{\chi}_{\rm ave}) \begin{bmatrix}R_{\rm tr} & (1+\rho)b_{\rm tr}\end{bmatrix}^\top\right) = \mathcal{O}_{\prec}(d^{-2}), \tag{111}$$

and thus we can simplify (108) as

$$\widehat{\chi}_{\rm ave} \simeq \frac{1}{d}\,\mathrm{tr}\left[\left(\frac{\tau}{1+\widehat{\chi}_{\rm ave}}E_{\rm tr} + \pi B + \lambda\tau I_d\right)^{-1}E_{\rm tr}\right]. \tag{112}$$

Observe that (112) is a small perturbation of the self-consistent equation in (76). By the stability of the equation (76), we then have

$$\widehat{\chi}_{\rm ave} \simeq \chi_\pi, \tag{113}$$

where χ_π is the unique positive solution to (76).

Recall the definitions of $\mathcal{G}_e(\pi)$ and $\widehat{\mathcal{G}}_e(\pi)$ in (100) and (77), respectively. By the standard resolvent identity,

$$\begin{aligned}&\widehat{\mathcal{G}}_e(\pi) - \mathcal{G}_e(\pi)\\ &= \frac{\tau[\widehat{\chi}_{\rm ave} - \chi_\pi]}{[1+\chi_\pi][1+\widehat{\chi}_{\rm ave}]}\widehat{\mathcal{G}}_e(\pi)\begin{bmatrix}1+\rho & \frac{1}{\sqrt{d}}\,\mathrm{vec}\left(\begin{bmatrix}R_{\rm tr} & (1+\rho)b_{\rm tr}\end{bmatrix}\right)^\top\\ \frac{1}{\sqrt{d}}\,\mathrm{vec}\left(\begin{bmatrix}R_{\rm tr} & (1+\rho)b_{\rm tr}\end{bmatrix}\right) & I_d\otimes E_{\rm tr}\end{bmatrix}\mathcal{G}_e(\pi).\end{aligned} \tag{114}$$

By construction, $\left\|\widehat{\mathcal{G}}_e(\pi)\right\|_{\mathsf{op}} \le 1/(\tau\lambda)$ and $\left\|\mathcal{G}_e(\pi)\right\|_{\mathsf{op}} \le 1/(\tau\lambda)$. Moreover, $\|E_{\rm tr}\|_{\mathsf{op}} \prec 1$ and

$$\left\|\frac{1}{\sqrt{d}}\,\mathrm{vec}\left(\begin{bmatrix}R_{\rm tr} & (1+\rho)b_{\rm tr}\end{bmatrix}\right)\right\| \prec 1. \tag{115}$$

It then follows from (113) and (114) that

$$\left\|\widehat{\mathcal{G}}_e(\pi) - \mathcal{G}_e(\pi)\right\|_{\mathsf{op}} \simeq 0. \tag{116}$$

If $\widehat{\mathcal{G}}_e(\pi)$ satisfies the equivalent conditions (73), (74) and (75) (as claimed in our analysis above), then the estimate in (116) allows us to easily check that $\mathcal{G}_e(\pi)$ also satisfies (73), (74) and (75). Thus, we claim that $\mathcal{G}_e(\pi)$ is asymptotically equivalent to the extended resolvent matrix $G_e(\pi)$ in the sense of Definition 1.

SI-13 Asymptotic Limits of the Generalization Errors

In this section, we use the characterization in Result 3 to derive the asymptotic limits of the generalization errors of associated with the set of parameters Γ^* learned from ridge regression.

SI-13.1 Asymptotic Limits of the Linear and Quadratic Terms

From Corollary 1 and the discussions in Remark 2, characterizing the test error $e(\Gamma^*)$ boils down to computing the linear term $\frac{1}{d}\operatorname{tr}\left(\Gamma^* A_{\text{test}}^\top\right)$ and the quadratic term $\frac{1}{d}\operatorname{tr}\left(\Gamma^* B_{\text{test}}(\Gamma^*)^\top\right)$, where A_{test} and B_{test} are the matrices defined in (41) and (42), respectively.

We consider test data distributions $\mathcal{P}_{\text{test}}$ as follows. From (35), the ICL task test setting we consider corresponds to choosing

$$\text{(ICL)}: \qquad A_{\text{test}} = \begin{bmatrix} I_d & 0 \end{bmatrix} \quad \text{and} \quad B_{\text{test}} = \begin{bmatrix} (\frac{1+\rho}{\alpha}+1)I_d & \\ & (1+\rho)^2 \end{bmatrix}. \tag{117}$$

.

Result 4. *Let Γ^* be the set of parameters learned from the ridge regression problem in (7). Let $A_{\text{test}} \in \mathbb{R}^{d\times(d+1)}$ and $B_{\text{test}} \in \mathbb{R}^{(d+1)\times(d+1)}$ be two matrices constructed as in (117). We have*

$$\frac{1}{d}\operatorname{tr}(\Gamma^* A_{\text{test}}^\top) \simeq \frac{1}{d}\operatorname{tr}\left(\Gamma_{\text{eq}}^* A_{\text{test}}^\top\right), \tag{118}$$

and

$$\frac{1}{d}\operatorname{tr}(\Gamma^* B_{\text{test}}(\Gamma^*)^\top) \simeq \frac{1}{d}\operatorname{tr}(\Gamma_{\text{eq}}^* B_{\text{test}}(\Gamma_{\text{eq}}^*)^T) - \frac{c_e}{d}\operatorname{tr}\left(B_{\text{test}}\left[(E_{\text{tr}}+\xi I)^{-1} - \xi(E_{\text{tr}}+\xi I)^{-2}\right]\right). \tag{119}$$

In the above displays, Γ_{eq}^* is an asymptotic equivalent of Γ^*, defined as

$$\Gamma_{\text{eq}}^* := \begin{bmatrix} R_{\text{tr}} & (1+\rho)b_{\text{tr}} \end{bmatrix}(E_{\text{tr}}+\xi I)^{-1}, \tag{120}$$

where ξ is the unique positive solution to the self-consistent equation

$$\xi\mathcal{M}_\kappa\left(\frac{1+\rho}{\alpha}+\xi\right) - \frac{\tau\lambda}{\xi} = 1-\tau, \tag{121}$$

and $\mathcal{M}_\kappa(\cdot)$ is the function defined in (177). Moreover, the scalar c_e in (119) is defined as

$$c_e = \frac{\rho+\nu-\nu^2\mathcal{M}_\kappa(\nu) - \xi\left[1-2\nu\mathcal{M}_\kappa(\nu)-\nu^2\mathcal{M}_\kappa'(\nu)\right]}{1-2\xi\mathcal{M}_\kappa(\nu)-\xi^2\mathcal{M}_\kappa'(\nu)-\tau}, \tag{122}$$

where

$$\nu := \frac{1+\rho}{\alpha}+\xi. \tag{123}$$

To derive the asymptotic characterizations (118) and (119) in Result 4, we first use block-matrix inversion to rewrite $\mathcal{G}_e(\pi)$ in (77) as

$$\mathcal{G}_e(\pi) = \begin{bmatrix} c^*(\pi) & -c^*(\pi)\left(q^*(\pi)\right)^\top \\ -c^*(\pi)\,q^*(\pi) & I\otimes F_E(\chi_\pi) + c^*(\pi)q^*(\pi)(q^*(\pi))^\top \end{bmatrix}, \tag{124}$$

where $F_E(\cdot)$ is the matrix-valued function defined in (105), *i.e.*,

$$F_E(\chi_\pi) = \left(\frac{\tau}{1+\chi_\pi}E_{\text{tr}} + \pi B_{\text{test}} + \lambda\tau I_{d+1}\right)^{-1}. \tag{125}$$

Moreover,

$$q^*(\pi) = \frac{\tau}{(1+\chi_\pi)\sqrt{d}}\operatorname{vec}\left(\begin{bmatrix} R_{\text{tr}} & (1+\rho)b_{\text{tr}} \end{bmatrix}F_E(\chi_\pi)\right), \tag{126}$$

and

$$\frac{1}{c^*(\pi)} = \frac{\tau(1+\rho)}{1+\chi_\pi} + \lambda\tau - \frac{\tau^2}{(1+\chi_\pi)^2 d}\operatorname{tr}\left(\begin{bmatrix} R_{\text{tr}} & (1+\rho)b_{\text{tr}} \end{bmatrix}F_E(\chi_\pi)\begin{bmatrix} R_{\text{tr}} & (1+\rho)b_{\text{tr}} \end{bmatrix}^\top\right). \tag{127}$$

Observe that there is a one-to-one correspondence between the terms in (124) and those in (57).

To derive the asymptotic characterization given in (118), we note that

$$\frac{1}{d}\operatorname{tr}(\Gamma^* A_{\text{test}}^\top) \simeq \frac{-1}{c(0)\sqrt{d}}\begin{bmatrix}0 & \operatorname{vec}(A_{\text{test}})^T\end{bmatrix}\mathcal{G}_e(0)e_1 \tag{128}$$

$$= \frac{c^*(0)}{c(0)}\cdot\frac{1}{d}\operatorname{tr}\left(\begin{bmatrix}R_{\text{tr}} & (1+\rho)b_{\text{tr}}\end{bmatrix}(E_{\text{tr}}+\lambda(1+\chi_0)I)^{-1}A_{\text{test}}^\top\right) \tag{129}$$

$$\simeq \frac{1}{d}\operatorname{tr}\left(\begin{bmatrix}R_{\text{tr}} & (1+\rho)b_{\text{tr}}\end{bmatrix}(E_{\text{tr}}+\lambda(1+\chi_0)I)^{-1}A_{\text{test}}^\top\right). \tag{130}$$

In the above display, (128) follows from (61) and the asymptotic equivalence between $G_e(0)$ and $\mathcal{G}_e(0)$. The equality in (129) is due to (124) and (126). To reach (130), we note that $c(0) = e_1^\top G_e(0)e_1$ and $c^*(0) = e_1^\top \mathcal{G}_e(0)e_1$. Thus, $c(0) \simeq c^*(0)$ due to the asymptotic equivalence between $G_e(0)$ and $\mathcal{G}_e(0)$. It can be shown that

$$\lambda(1+\chi_0) \simeq \xi, \tag{131}$$

where ξ is the scalar defined in (121). The asymptotic characterization given in (118) then follows from (130) and from the definition of Γ_{eq}^* given in (120).

Next, we use (63) to derive the asymptotic characterization of the quadratic term in (119). Taking the derivative of (127) gives us

$$\frac{\mathrm{d}}{\mathrm{d}\pi}\left(\frac{1}{c^*(\pi)}\right)\Big|_{\pi=0} = \frac{1}{d}\operatorname{tr}(\Gamma_{\text{eq}}^* B_{\text{test}}(\Gamma_{\text{eq}}^*)^\top) - \frac{\tau\chi_0'}{(1+\chi_0)^2}\left(1+\rho-\frac{2}{d}\operatorname{tr}(A_{\text{tr}}(E_{\text{tr}}+\xi I)^{-1}A_{\text{tr}}^T)+\frac{1}{d}\operatorname{tr}(A_{\text{tr}}(E_{\text{tr}}+\xi I)^{-1}E_{\text{tr}}(E_{\text{tr}}+\xi I)^{-1}A_{\text{tr}}^T)\right) \tag{132}$$

$$= \frac{1}{d}\operatorname{tr}(\Gamma_{\text{eq}}^* B_{\text{test}}(\Gamma_{\text{eq}}^*)^\top) - \frac{\tau\chi_0'}{(1+\chi_0)^2}\left(1+\rho-\frac{1}{d}\operatorname{tr}(\Gamma_{\text{eq}}^* A_{\text{tr}}^T)-\frac{\xi}{d}\operatorname{tr}(\Gamma_{\text{eq}}^*(\Gamma_{\text{eq}}^*)^\top)\right), \tag{133}$$

where A_{tr} is the matrix defined in (71). In reaching the above expression, we have also used the estimate in (131).

To further simplify our formula, we note that

$$A_{\text{tr}} = S\left(E_{\text{tr}}+\xi I_{d+1}-\left(\frac{1+\rho}{\alpha}+\xi\right)I_{d+1}\right), \tag{134}$$

where S is a $d\times(d+1)$ matrix obtained by removing the last row of I_{d+1}. Using this identity, we can rewrite the matrix Γ_{eq}^* in (120) as

$$\Gamma_{\text{eq}}^* = S\left(I-\left(\frac{1+\rho}{\alpha}+\xi\right)(E_{\text{tr}}+\xi I)^{-1}\right) \tag{135}$$

$$= \begin{bmatrix}I-\nu F_R(\nu)-a^*(1+\rho)^2\nu F_R(\nu)b_{\text{tr}}b_{\text{tr}}^\top F_R(\nu) & a^*(1+\rho)\nu F_R(\nu)b_{\text{tr}}\end{bmatrix}, \tag{136}$$

where $F_R(\cdot)$ is the function defined in (175), and ν is the parameter given in (123). The second equality (136) is obtained from the explicit formula for $(E_{\text{tr}}+\xi I)^{-1}$ in (179).

From (134) and (135), it is straightforward to check that

$$\frac{1}{d}\operatorname{tr}(\Gamma_{\text{eq}}^* A_{\text{tr}}^T) = 1-\nu+\nu^2\frac{1}{d}\operatorname{tr}(S(E_{\text{tr}}+\xi I)^{-1}S^\top), \tag{137}$$

and

$$\frac{\xi}{d}\operatorname{tr}(\Gamma_{\text{eq}}^*(\Gamma_{\text{eq}}^*)^\top) = \xi\left[1-2\nu\frac{1}{d}\operatorname{tr}(S(E_{\text{tr}}+\xi I)^{-1}S^\top)+\nu^2\frac{1}{d}\operatorname{tr}(S(E_{\text{tr}}+\xi I)^{-2}S^\top\right]. \tag{138}$$

By using the asymptotic characterizations given in (182) and (183), we then have

$$\frac{1}{d}\operatorname{tr}(\Gamma_{\text{eq}}^* A_{\text{tr}}^T) \simeq 1-\nu+\nu^2\mathcal{M}_\kappa(\nu), \tag{139}$$

and

$$\frac{\xi}{d}\operatorname{tr}(\Gamma_{\text{eq}}^*(\Gamma_{\text{eq}}^*)^\top) \simeq \xi\left[1-2\nu\mathcal{M}_\kappa(\nu)-\nu^2\mathcal{M}_\kappa'(\nu)\right]. \tag{140}$$

Substituting (139), (140), and (184) into (133) yields

$$\frac{\mathrm{d}}{\mathrm{d}\pi}\left(\frac{1}{c^*(\pi)}\right)\Big|_{\pi=0} \simeq \frac{1}{d}\operatorname{tr}(\Gamma_{\mathrm{eq}}^* B_{\mathrm{test}}(\Gamma_{\mathrm{eq}}^*)^T) - \frac{c_e}{d}\operatorname{tr}\left(B_{\mathrm{test}}\left[(E_{\mathrm{tr}}+\xi I)^{-1} - \xi(E_{\mathrm{tr}}+\xi I)^{-2}\right]\right), \tag{141}$$

where c_e is the scalar defined in (122). The asymptotic characterization of the quadratic term in (119) then follows from (63) and the claim that

$$\frac{\mathrm{d}}{\mathrm{d}\pi}\left(\frac{1}{c(\pi)}\right)\Big|_{\pi=0} \simeq \frac{\mathrm{d}}{\mathrm{d}\pi}\left(\frac{1}{c^*(\pi)}\right)\Big|_{\pi=0}. \tag{142}$$

SI-13.2 The Generalization Error of In-Context Learning

Result 5. *Consider the test distribution $\mathcal{P}_{\mathrm{test}}$ associated with the ICL task. We have*

$$e(\Gamma^*) \simeq e^{\mathrm{ICL}}(\tau,\alpha,\kappa,\rho,\lambda), \tag{143}$$

where

$$\begin{aligned} e^{\mathrm{ICL}}(\tau,\alpha,\kappa,\rho,\lambda) \coloneqq \left(\frac{1+\rho}{\alpha}+1\right)\left(1-2\nu\mathcal{M}_\kappa(\nu)-\nu^2\mathcal{M}_\kappa'(\nu)-c_e\left[\mathcal{M}_\kappa(\nu)+\xi\mathcal{M}_\kappa'(\nu)\right]\right) \\ -2\left[1-\nu\mathcal{M}_\kappa(\nu)\right]+1+\rho, \end{aligned} \tag{144}$$

and c_e is the constant given in (122).

Remark 6. *Recall the definition of the asymptotic equivalence notation "$\simeq$" introduced in Section SI-9. The characterization given in (143) implies that, as $d\to\infty$, the generalization error $e(\Gamma^*)$ converges almost surely to the deterministic quantity $e^{\mathrm{ICL}}(\tau,\alpha,\kappa,\rho,\lambda)$.*

To derive (143), our starting point is the estimate

$$e(\Gamma^*) \simeq \frac{1}{d}\operatorname{tr}\left(\Gamma^* B_{\mathrm{test}}(\Gamma^*)^\top\right) - \frac{2}{d}\operatorname{tr}\left(\Gamma^* A_{\mathrm{test}}^\top\right) + 1 + \rho, \tag{145}$$

which follows from Corollary 1 and the discussions in Remark 2. We consider the ICL task here, and thus A_{test} and B_{test} are given in (117). The asymptotic limits of the first two terms on the right-hand side of the above equation can be obtained by the characterizations given in Result 4.

Using (118) and the expressions in (136) and (117), we have

$$\frac{1}{d}\operatorname{tr}(\Gamma^* A_{\mathrm{test}}^\top) \simeq \frac{1}{d}\operatorname{tr}\left(\Gamma_{\mathrm{eq}}^* A_{\mathrm{test}}^\top\right) \tag{146}$$

$$= 1 - \frac{\nu}{d}\operatorname{tr} F_R(\nu) - a^*(1+\rho)^2\nu\frac{\left\|F_R(\nu)b_{\mathrm{tr}}\right\|^2}{d} \tag{147}$$

$$\simeq 1-\nu\mathcal{M}_\kappa(\nu), \tag{148}$$

where ν is the constant defined in (123). To reach the last step, we have used the estimate given in (182).

Next, we use (119) to characterize the first term on the right-hand side of (145). From the formulas in (136) and (117), we can check that

$$\frac{1}{d}\operatorname{tr}\left(\Gamma_{\mathrm{eq}}^* B_{\mathrm{test}}(\Gamma_{\mathrm{eq}}^*)^\top\right) \simeq \left(\frac{1+\rho}{\alpha}+1\right)\frac{1}{d}\operatorname{tr}\left(I-\nu F(\nu)\right)^2 \tag{149}$$

$$\simeq \left(\frac{1+\rho}{\alpha}+1\right)\left(1-2\nu\mathcal{M}_\kappa(\nu)-\nu^2\mathcal{M}_\kappa'(\nu)\right), \tag{150}$$

where the second step follows from (182) and (183). From (179),

$$\frac{1}{d}\operatorname{tr}(B_{\mathrm{test}}(E_{\mathrm{tr}}+\xi I)^{-1}) \simeq \left(\frac{1+\rho}{\alpha}+1\right)\frac{1}{d}\operatorname{tr} F_R(\nu) \simeq \left(\frac{1+\rho}{\alpha}+1\right)\mathcal{M}_\kappa(\nu). \tag{151}$$

Similarly, we can check that

$$\frac{1}{d}\operatorname{tr}(B_{\mathrm{test}}(E_{\mathrm{tr}}+\xi I)^{-2}) \simeq \left(\frac{1+\rho}{\alpha}+1\right)\frac{1}{d}\operatorname{tr} F_R^2(\nu) \simeq -\left(\frac{1+\rho}{\alpha}+1\right)\mathcal{M}_\kappa'(\nu). \tag{152}$$

Substituting (150), (151), and (152) into (119) gives us

$$\frac{1}{d}\operatorname{tr}(\Gamma^* B(\Gamma^*)^\top) \simeq \left(\frac{1+\rho}{\alpha}+1\right)\left(1-2\nu\mathcal{M}_\kappa(\nu)-\nu^2\mathcal{M}_\kappa'(\nu)-c_e\left[\mathcal{M}_\kappa(\nu)+\xi\mathcal{M}_\kappa'(\nu)\right]\right), \tag{153}$$

where c_e is the constant given in (122). Combining (148), (153), and (145), we are done.

In what follows, we further simplify the characterizations in Result 5 by considering the ridgeless limit, *i.e.*, when $\lambda \to 0^+$.

Result 6. *Let*

$$q^* \coloneqq \frac{1+\rho}{\alpha}, \qquad m^* \coloneqq \mathcal{M}_\kappa\left(q^*\right), \qquad \textit{and} \qquad \mu^* \coloneqq q^*\mathcal{M}_{\kappa/\tau}(q^*), \tag{154}$$

where $\mathcal{M}_\kappa(x)$ is the function defined in (177). Then

$$\begin{aligned} e^{\mathrm{ICL}}_{\mathrm{ridgeless}} &\coloneqq \lim_{\lambda\to 0^+} e^{\mathrm{ICL}}(\tau,\alpha,\kappa,\rho,\lambda) \\ &= \begin{cases} \frac{\tau(1+q^*)}{1-\tau}\left[1-\tau(1-\mu^*)^2+\mu^*(\rho/q^*-1)\right]-2\tau(1-\mu^*)+(1+\rho) & \tau<1 \\ (q^*+1)\left(1-2q^*m^*-(q^*)^2\mathcal{M}_\kappa'(q^*)+\frac{(\rho+q^*-(q^*)^2m^*)m^*}{\tau-1}\right)-2(1-q^*m^*)+(1+\rho) & \tau>1 \end{cases}, \end{aligned} \tag{155}$$

where $\mathcal{M}_\kappa'(\cdot)$ denotes the derivative of $\mathcal{M}_\kappa(x)$ with respect to x.

We start with the case of $\tau < 1$. Examining the self-consistent equation in (121), we can see that the parameter ξ tends to a nonzero constant, denoted by ξ^*, as $\lambda \to 0^+$. It follows that the original equation in (121) reduces to

$$\xi^*\mathcal{M}_\kappa\left(\frac{1+\rho}{\alpha}+\xi^*\right) = 1-\tau. \tag{156}$$

Introduce a change of variables

$$\mu^* \coloneqq \frac{(1-\tau)(1+\rho)}{\alpha\tau\xi^*}. \tag{157}$$

By combining (156) and the characterization in (178), we can directly solve for μ and get $\mu^* = q^*\mathcal{M}_{\kappa/\tau}(q^*)$ as given in (154). The characterization in (155) (for the case of $\tau < 1$) then directly follows from (148), (153), and (4) after some lengthy calculations.

Next, we consider the case of $\tau > 1$. It is straightforward to verify from (121) that

$$\xi = \frac{\tau}{\tau-1}\lambda + \mathcal{O}(\lambda^2). \tag{158}$$

Thus, when $\tau > 1$, $\xi \to 0$ as $\lambda \to 0^+$. It follows that

$$\lim_{\lambda\to 0^+}\nu = \lim_{\lambda\to 0^+}\left(\frac{1+\rho}{\alpha}+\xi\right) = q^* \quad \text{and} \quad \lim_{\lambda\to 0^+}\mathcal{M}_\kappa(\nu) = m^*. \tag{159}$$

Substituting these estimates into (148), (153), and (4), we then reach the characterizations in (155) for the case of $\tau > 1$.

SI-13.3 The Generalization Error of In-Distribution Generalization

In what follows, we derive the asymptotic limit of the generalization error for the IDG task.

Result 7. *Consider the test distribution $\mathcal{P}_{\mathrm{test}}$ associated with the IDG task. We have*

$$e(\Gamma^*) \simeq e^{\mathrm{IDG}}(\tau,\alpha,\kappa,\rho,\lambda) \coloneqq \tau\frac{\rho+\nu-\nu^2\mathcal{M}_\kappa(\nu)-\xi\left[1-2\nu\mathcal{M}_\kappa(\nu)-\nu^2\mathcal{M}_\kappa'(\nu)\right]}{\tau-\left[1-2\xi\mathcal{M}_\kappa(\nu)-\xi^2\mathcal{M}_\kappa'(\nu)\right]}, \tag{160}$$

where ξ the unique positive solution to the self-consistent equation (121) and ν is the constant given in (123).

Similar to our derivation of Result 5, we only need to use (118) and (119) to characterize the asymptotic limits of the first and second terms on the right-hand side of (145). Note that, for the IDG task, $A_{\text{test}} = A_{\text{tr}}$. It follows from (118) and (139) that

$$\frac{1}{d}\operatorname{tr}(\Gamma^* A_{\text{test}}^\top) \simeq 1 - \nu + \nu^2 \mathcal{M}_\kappa(\nu). \tag{161}$$

Similarly, since $B_{\text{test}} = E_{\text{tr}}$, we can verify from (120) that

$$\frac{1}{d}\operatorname{tr}\left(\Gamma_{\text{eq}}^* B_{\text{test}} (\Gamma_{\text{eq}}^*)^\top\right) = \frac{1}{d}\operatorname{tr}(\Gamma_{\text{eq}}^* A_{\text{tr}}^\top) - \frac{\xi}{d}\operatorname{tr}(\Gamma_{\text{eq}}^* (\Gamma_{\text{eq}}^*)^\top) \tag{162}$$

$$\simeq 1 - \nu + \nu^2 \mathcal{M}_\kappa(\nu) - \xi\left[1 - 2\nu\mathcal{M}_\kappa(\nu) - \nu^2 \mathcal{M}_\kappa'(\nu)\right], \tag{163}$$

where the second step follows from (139) and (140). Moreover,

$$\frac{1}{d}\operatorname{tr}\left(B_{\text{test}}\left[(E_{\text{tr}} + \xi I)^{-1} - \xi(E_{\text{tr}} + \xi I)^{-2}\right]\right) = 1 - 2\xi\mathcal{M}_\kappa(\nu) - \xi^2 \mathcal{M}_\kappa'(\nu). \tag{164}$$

Substituting (162) and (164) into (119), we have

$$\begin{aligned}
&\frac{1}{d}\operatorname{tr}(\Gamma^* B (\Gamma^*)^\top) \\
&\simeq \tau \frac{\rho + \nu - \nu^2\mathcal{M}_\kappa(\nu) - \xi\left[1 - 2\nu\mathcal{M}_\kappa(\nu) - \nu^2\mathcal{M}_\kappa'(\nu)\right]}{\tau - \left[1 - 2\xi\mathcal{M}_\kappa(\nu) - \xi^2\mathcal{M}_\kappa'(\nu)\right]} + 2(1 - \nu + \nu^2\mathcal{M}_\kappa(\nu)) - (1+\rho).
\end{aligned} \tag{165}$$

The final result in (160) then follows from combining the above expression with (161) and (145).

Finally, we derive the ridgeless limit of the characterization given in Result 7.

Result 8. *Let q^*, m^*, and μ^* be the scalars defined in (154). We have*

$$e_{\text{ridgeless}}^{\text{IDG}} := \lim_{\lambda \to 0^+} e^{\text{IDG}}(\tau, \alpha, \kappa, \rho, \lambda) \tag{166}$$

$$= \begin{cases} \frac{\tau}{1-\tau}\left(\frac{\rho + q^* - 2q^*(1-\tau)(q^*/\xi^* + 1)}{1 - p^*(1-\tau)} + \frac{\tau\mu^*(q^* + \xi^*)^2}{q^*}\right) & \tau < 1 \\ \frac{\tau}{\tau - 1}[\rho + q^*(1 - q^* m^*)] & \tau > 1 \end{cases}, \tag{167}$$

where $\xi^* = \frac{(1-\tau)q^*}{\tau\mu^*}$ and $p^* = \left(1 - \kappa\left(\frac{\kappa\xi^*}{1-\tau} + 1\right)^{-2}\right)^{-1}$.

The derivation of this result closely follows that of Result 6. We analyze the cases of $\tau < 1$ and $\tau > 1$ separately. For $\tau < 1$, the equation in (121) simplifies to (156) as $\lambda \to 0^+$. For $\tau > 1$, ξ approaches zero as $\lambda \to 0^+$. Substituting these estimates into (160) then yields (167) after some detailed calculations.

SI-14 Asymptotics of dMMSE Estimator

To study the slowness of the dMMSE estimator more explicitly, consider the $\alpha \to \infty$ limit. We present an initial sketch for the rate at which $g_{\text{task}} \to 0$ in this limit, by considering large k for d fixed.

The exponential weight terms in the estimator in this limit behave as

$$e^{-\frac{1}{2\rho}\sum_{i=1}^{\ell}\left(y_i - w_j^\top x_i\right)^2} \to e^{-\frac{\ell}{2\rho}\left(\frac{1}{d}\|w^* - w_j\|^2 + \rho\right)},$$

and these weightings exponentially favor choosing ω_j that minimises $\|w^* - w_j\|^2$ over the set of k training tasks w_j. It's immediately clear that $e_{\text{IDG}}^{\text{dMMSE}} = \rho$ in this limit as the minimal value of $\|w^* - w_j\|^2$ when $w^* \in \{w_1, \cdots, w_k\}$ is 0. Taking

$$w_{\text{est}}(w^*, w_i) = \text{argmin}_{i \in [k]} \|w^* - w_i\|^2$$

we have

$$g_{\text{task}}^{\text{dMMSE}} = e_{\text{ICL}}^{\text{dMMSE}} - \rho = \frac{1}{d}\mathbb{E}_{w^* \sim \mathcal{P}_{\text{test}}}\left[\mathbb{E}_{w_i \sim \mathcal{P}_{\text{train}}}\left[\min_{i \in [k]} \|w^* - w_i\|^2\right]\right] \tag{168}$$

We exploit spherical symmetry in both w^* and w_i to simplify

$$\|w^* - w_i\|^2 = 4d - 4db_i \qquad \text{for } b_i \sim \text{Beta}\left(\frac{d-1}{2}, \frac{d-1}{2}\right).$$

A rate of convergence can then be derived by studying the asymptotic behaviour of the expected maximum of the Beta distribution. This is an ongoing computation, with results suggesting exponentially-cursed decay in dimension. Specifically, heuristic arguments integrating over the PDF of $\max_i b_i$ suggests $\mathcal{O}(\kappa^{-2/d})$ decay.

A Auxiliary Results

Lemma 3. *Let w be a given task vector with $\|w\| = \sqrt{d}$. Meanwhile, let $a \sim \mathcal{N}(0,1)$, $s \sim \mathcal{N}(0,1)$, $\epsilon \sim \mathcal{N}(0,\rho)$ be three scalar normal random variables, and $q \sim \mathcal{N}(0, I_{\ell-1})$, $g \sim \mathcal{N}(0, I_{d-1})$, $u \sim \mathcal{N}(0, I_{d-1})$, and $v_\epsilon \sim \mathcal{N}(0, \rho I_\ell)$ be isotropic normal random vectors. Moreover, w and all of the above random variables are mutually independent. We have the following equivalent statistical representation of the pair $(H_Z, y_{\ell+1})$:*

$$H_Z \overset{(d)}{=} (d/\ell) M_w \begin{bmatrix} s \\ u \end{bmatrix} \left[h^\top M_w, \quad (a/\sqrt{d} + \theta_\epsilon)^2/\sqrt{d} + \theta_q^2/\sqrt{d} \right], \tag{169}$$

and

$$y_{\ell+1} \overset{(d)}{=} s + \epsilon. \tag{170}$$

In the above displays, M_w denotes a symmetric *and* orthonormal *matrix such that*

$$(M_w) e_1 = \frac{w}{\|w\|}, \tag{171}$$

where e_1 denotes the first natural basis vector in $\mathbb{R}^d$; $h \in \mathbb{R}^d$ is a vector defined as

$$h \coloneqq \begin{bmatrix} \frac{\theta_\epsilon a}{\sqrt{d}} + \frac{a^2}{d} + \theta_q^2 \\ \left[(\theta_\epsilon + a/\sqrt{d})^2 + \theta_q^2\right]^{1/2} g/\sqrt{d} \end{bmatrix}; \tag{172}$$

and θ_ϵ, θ_q are scalars such that

$$\theta_\epsilon = \|v_\epsilon\|/\sqrt{d} \qquad \textit{and} \qquad \theta_q = \|q\|/\sqrt{d}. \tag{173}$$

We will also find it useful to note

$$X^\top w = M_{v_\epsilon} \begin{bmatrix} a \\ q \end{bmatrix}, \tag{174}$$

Define the following resolvent

$$F_R(\nu) \coloneqq (R_{\text{tr}} + \nu I_d)^{-1}, \tag{175}$$

where R_{tr} is the sample covariance matrix of the task vectors as defined in (70) and ν is a positive scalar.

Note that the distribution of R_{tr} is asymptotically equivalent to that of a Wishart ensemble. By standard random matrix results on the Stieltjes transforms of Wishart ensembles (see, *e.g.*, [29]), we have

$$\frac{1}{d} \operatorname{tr} F_R(\nu) \simeq \mathcal{M}_\kappa(\nu) \tag{176}$$

as $d, k \to \infty$ with $k/d = \kappa$. Here,

$$\mathcal{M}_\kappa(\nu) \coloneqq \frac{2}{\nu + 1 - 1/\kappa + \left[(\nu + 1 - 1/\kappa)^2 + 4\nu/\kappa\right]^{1/2}}. \tag{177}$$

is the solution to the self-consistent equation

$$\frac{1}{\mathcal{M}_\kappa(\nu)} = \frac{1}{1 + \mathcal{M}_\kappa(\nu)/\kappa} + \nu. \tag{178}$$

$$(E_{\rm tr}+\lambda(1+\chi_0)I_{d+1})^{-1} = \begin{bmatrix} F_R(\nu_0)+a^*(1+\rho)^2F_R(\nu_0)b_{\rm tr}b_{\rm tr}^\top F_R(\nu_0) & -a^*(1+\rho)F_R(\nu_0)b_{\rm tr} \\ -a^*(1+\rho)b_{\rm tr}^\top F_R(\nu_0) & a^* \end{bmatrix}, \tag{179}$$

where $F_R(\cdot)$ is the function defined in (175),

$$\nu_0 = \frac{1+\rho}{\alpha} + \lambda(1+\chi_0) \tag{180}$$

and

$$\frac{1}{a^*} = (1+\rho)^2 + \lambda(1+\chi_0) - (1+\rho)^2 b_{\rm tr}^\top F_R(\nu_0) b_{\rm tr}. \tag{181}$$

$$\frac{1}{d}\operatorname{tr}(E_{\rm tr}+\xi I)^{-1} \simeq \frac{1}{d}\operatorname{tr} S(E_{\rm tr}+\xi I)^{-1}S^\top \simeq \frac{1}{d}\operatorname{tr} F\left(\frac{1+\rho}{\alpha}+\xi\right) \simeq \mathcal{M}_\kappa\left(\frac{1+\rho}{\alpha}+\xi\right), \tag{182}$$

and

$$\frac{1}{d}\operatorname{tr}(E_{\rm tr}+\xi I)^{-2} \simeq \frac{1}{d}\operatorname{tr} S(E_{\rm tr}+\xi I)^{-2}S^\top \simeq \frac{1}{d}\operatorname{tr} F^2\left(\frac{1+\rho}{\alpha}+\xi\right) \simeq -\mathcal{M}'_\kappa\left(\frac{1+\rho}{\alpha}+\xi\right), \tag{183}$$

where S is a $d\times(d+1)$ matrix obtained by removing the last row of I_{d+1}, and $\mathcal{M}_\kappa(\cdot)$ is the function defined in (177). Upon substitution have

$$\frac{\tau\chi_0'}{(1+\chi_0)^2} \simeq \frac{\frac{1}{d}\operatorname{tr}\left(B_{\rm test}[(E_{\rm tr}+\xi I)^{-1} - \xi(E_{\rm tr}+\xi I)^{-2}]\right)}{1-2\xi\mathcal{M}_\kappa\left(\frac{1+\rho}{\alpha}+\xi\right) - \xi^2\mathcal{M}'_\kappa\left(\frac{1+\rho}{\alpha}+\xi\right) - \tau}. \tag{184}$$